Iceland
Liechtenstein
Norway grants

THESEUS
Connect the Disconnections -
from Disparate Data to Insightful Analysis

NTNU CCIS

Computational Thinking - Big Data Analysis Process

Katrin FRANKE, PhD Professor of Computer Science

Center for Cyber and Information Security | www.ccis.no
Norwegian University of Science and Technology | www.ntnu.no

Contact: www.ntnu.edu/employees/katrin.franke

This publication was realised with the EEA Financial Mechanism 2014-2021 financial support. Its content (text, photos, videos) does not reflect the official opinion of the Programme Operator, the National Contact Point and the Financial Mechanism Office. Responsibility for the information and views expressed therein lies entirely with the author(s).

NORWEGIAN UNIVERSITY OF SCIENCE AND TECHNOLOGY

KNOWLEDGE FOR A BETTER WORLD

TRONDHEIM | GJØVIK | ÅLESUND
BRUSSELS | TOKYO

NTNU

3

NTNU* Digital

- AI - Machine Learning
- Security & Reliability
- Computation
- Autonomous Systems

* Norwegian University of Science and Technology

4

This publication was realised with the EEA Financial Mechanism 2014-2021 financial support. Its content (text, photos, videos) does not reflect the official opinion of the Programme Operator, the National Contact Point and the Financial Mechanism Office. Responsibility for the information and views expressed therein lies entirely with the author(s).

NTNU - Knowledge for a better world

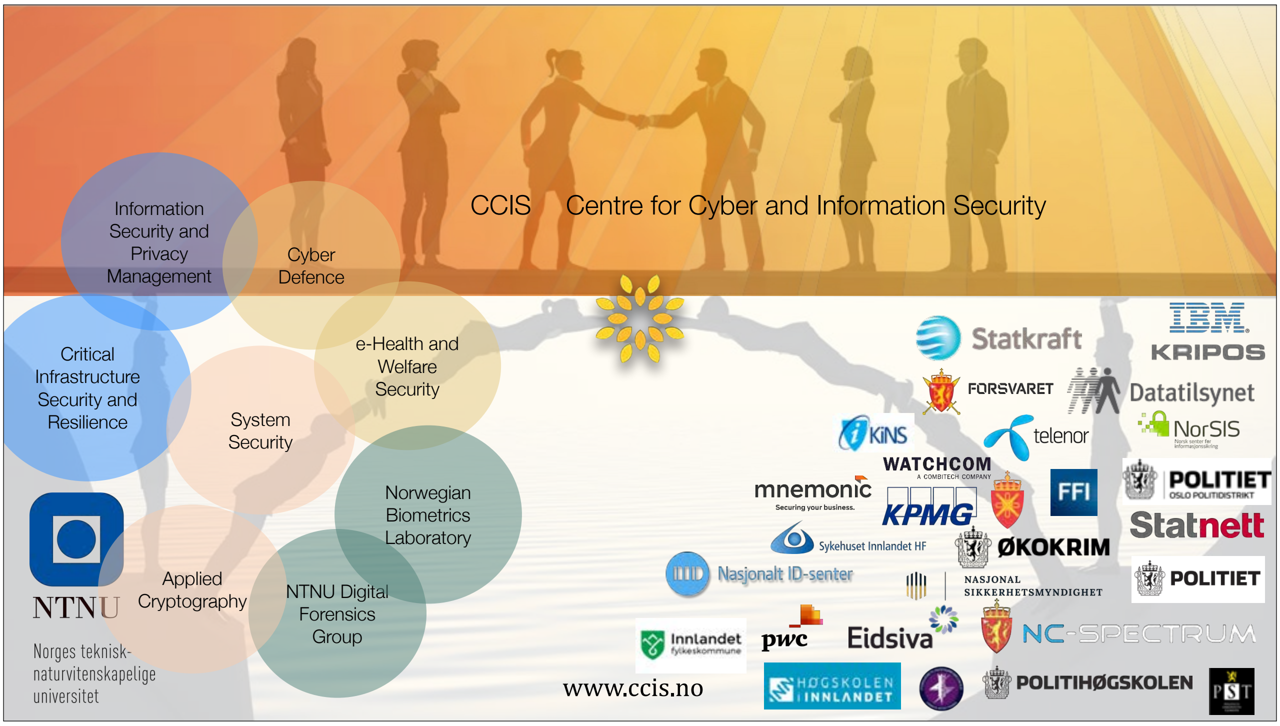
Our Vision

- Fighting ‘stupidity’ and evil
- Creating the machine that goes ‘Bing!’
- Finding and understanding the truth.

Security & Reliability
AI-Machine learning

* Norwegian University of Science and Technology 5

5



CCIS Centre for Cyber and Information Security

Information Security and Privacy Management

Cyber Defence

Critical Infrastructure Security and Resilience

e-Health and Welfare Security

System Security

Norwegian Biometrics Laboratory

Applied Cryptography

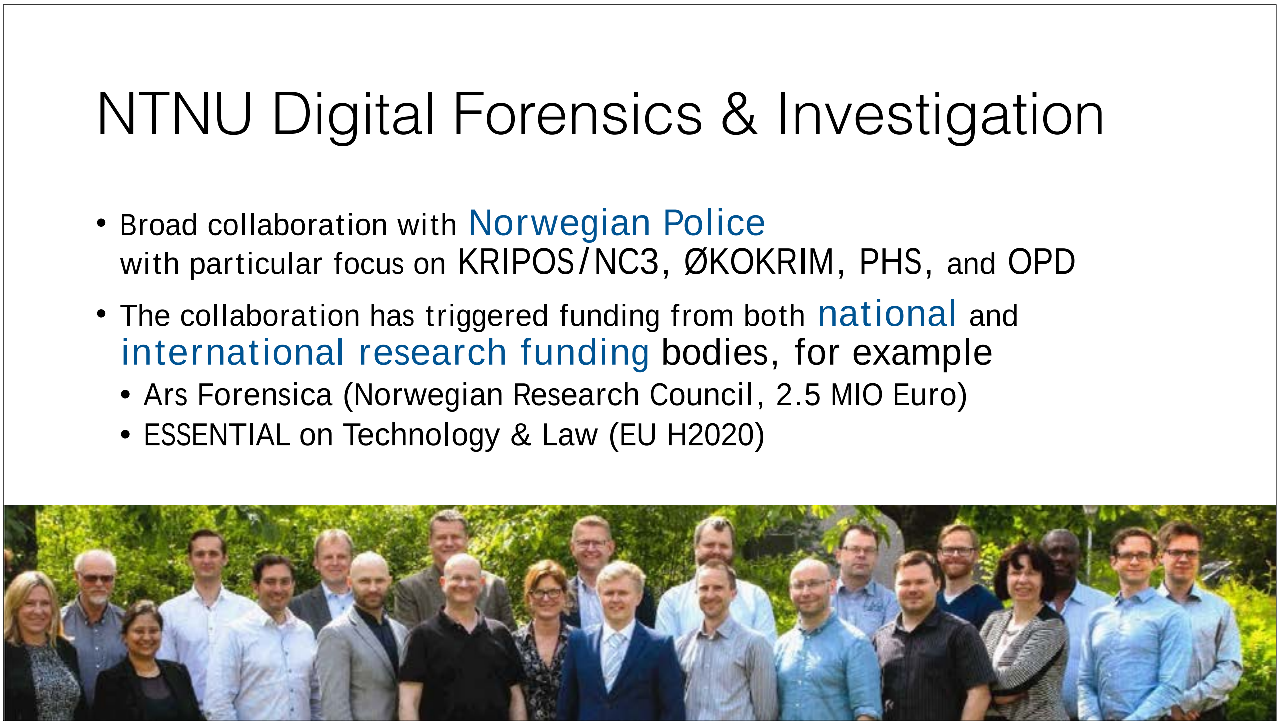
NTNU Digital Forensics Group

NTNU
Norges teknisk-naturvitenskapelige universitet

www.ccis.no

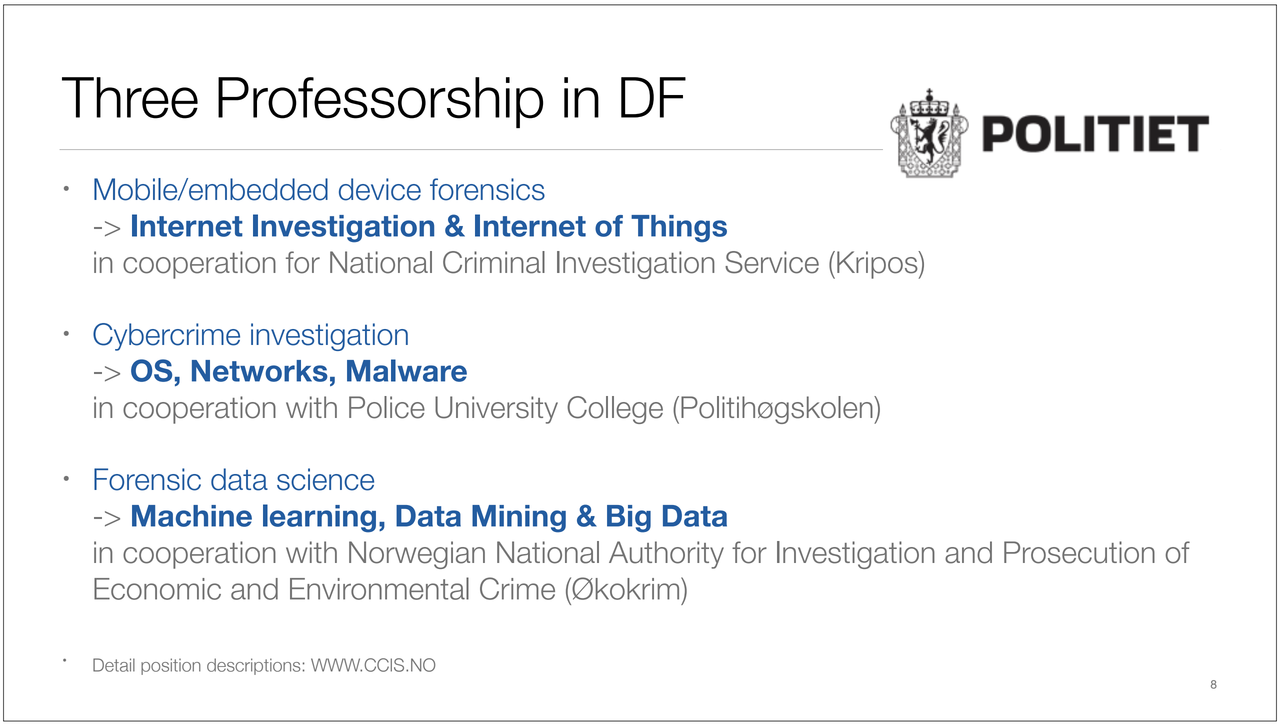
6

This publication was realised with the EEA Financial Mechanism 2014-2021 financial support. Its content (text, photos, videos) does not reflect the official opinion of the Programme Operator, the National Contact Point and the Financial Mechanism Office. Responsibility for the information and views expressed therein lies entirely with the author(s).



NTNU Digital Forensics & Investigation

- Broad collaboration with Norwegian Police with particular focus on KRIPOS/NC3, ØKOKRIM, PHS, and OPD
- The collaboration has triggered funding from both national and international research funding bodies, for example
  - Ars Forensica (Norwegian Research Council, 2.5 MIO Euro)
  - ESSENTIAL on Technology & Law (EU H2020)



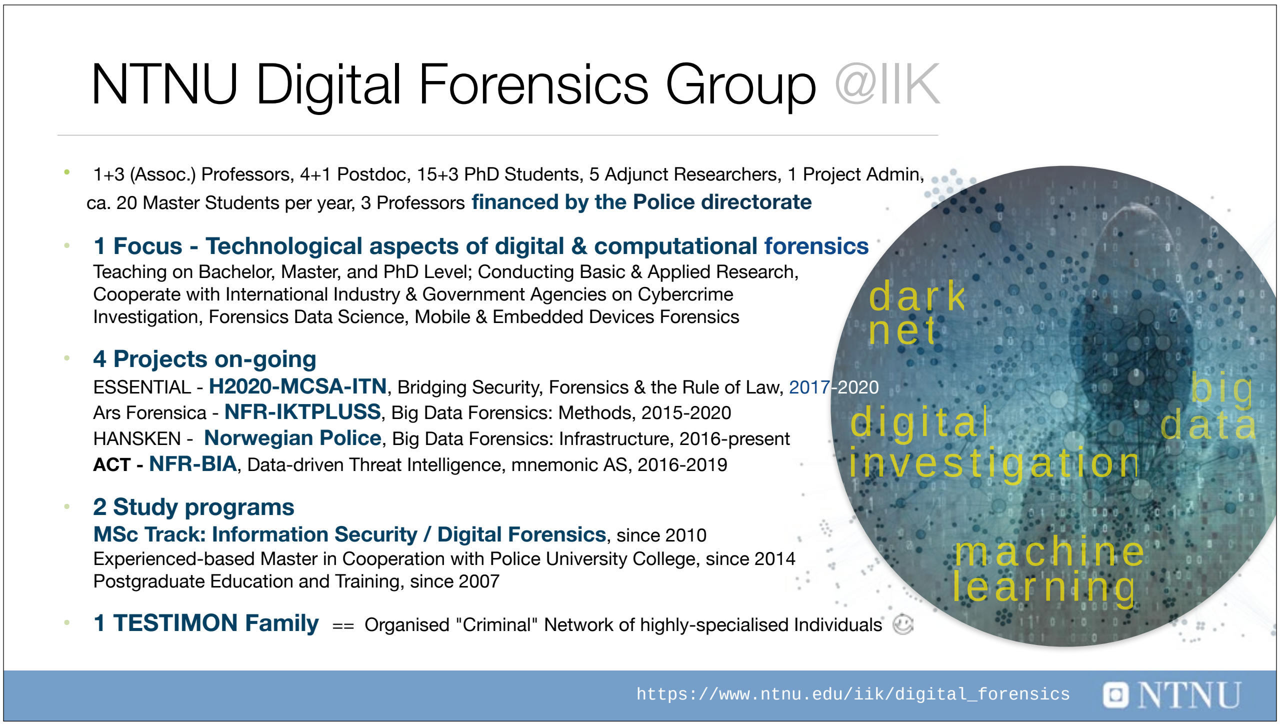
Three Professorship in DF

POLITIET

- Mobile/embedded device forensics
  -> **Internet Investigation & Internet of Things**
  in cooperation for National Criminal Investigation Service (Kripos)
- Cybercrime investigation
  -> **OS, Networks, Malware**
  in cooperation with Police University College (Politihøgskolen)
- Forensic data science
  -> **Machine learning, Data Mining & Big Data**
  in cooperation with Norwegian National Authority for Investigation and Prosecution of Economic and Environmental Crime (Økokrim)
- Detail position descriptions: WWW.CCIS.NO

8

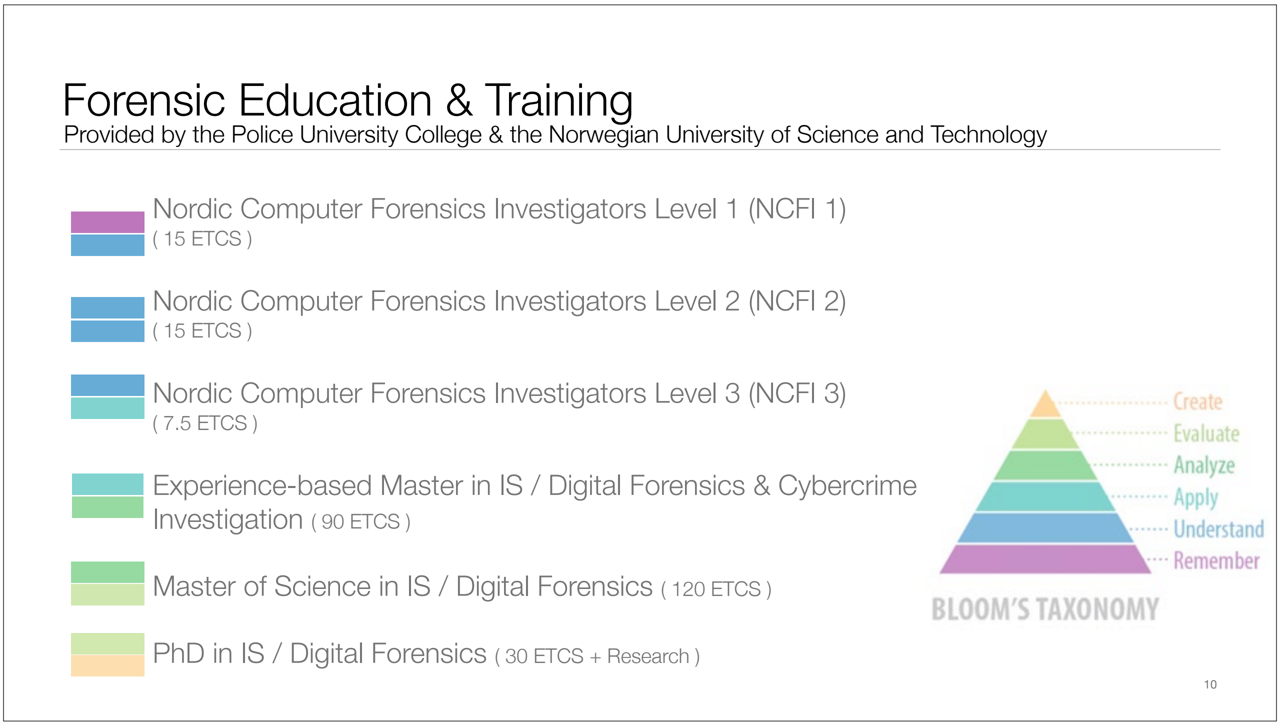
This publication was realised with the EEA Financial Mechanism 2014-2021 financial support. Its content (text, photos, videos) does not reflect the official opinion of the Programme Operator, the National Contact Point and the Financial Mechanism Office. Responsibility for the information and views expressed therein lies entirely with the author(s).



NTNU Digital Forensics Group @IIK

- 1+3 (Assoc.) Professors, 4+1 Postdoc, 15+3 PhD Students, 5 Adjunct Researchers, 1 Project Admin, ca. 20 Master Students per year, 3 Professors **financed by the Police directorate**
- **1 Focus - Technological aspects of digital & computational forensics**
  Teaching on Bachelor, Master, and PhD Level; Conducting Basic & Applied Research, Cooperate with International Industry & Government Agencies on Cybercrime Investigation, Forensics Data Science, Mobile & Embedded Devices Forensics
- **4 Projects on-going**
  ESSENTIAL - **H2020-MCSA-ITN**, Bridging Security, Forensics & the Rule of Law, 2017-2020
  Ars Forensica - **NFR-IKTPLUSS**, Big Data Forensics: Methods, 2015-2020
  HANSKEN - **Norwegian Police**, Big Data Forensics: Infrastructure, 2016-present
  ACT - **NFR-BIA**, Data-driven Threat Intelligence, mnemonic AS, 2016-2019
- **2 Study programs**
  MSc Track: Information Security / Digital Forensics, since 2010
  Experienced-based Master in Cooperation with Police University College, since 2014
  Postgraduate Education and Training, since 2007
- **1 TESTIMON Family** == Organised "Criminal" Network of highly-specialised Individuals ☺

https://www.ntnu.edu/iik/digital_forensics NTNU



Forensic Education & Training

Provided by the Police University College & the Norwegian University of Science and Technology

Nordic Computer Forensics Investigators Level 1 (NCFI 1)
(15 ETCS)

Nordic Computer Forensics Investigators Level 2 (NCFI 2)
(15 ETCS)

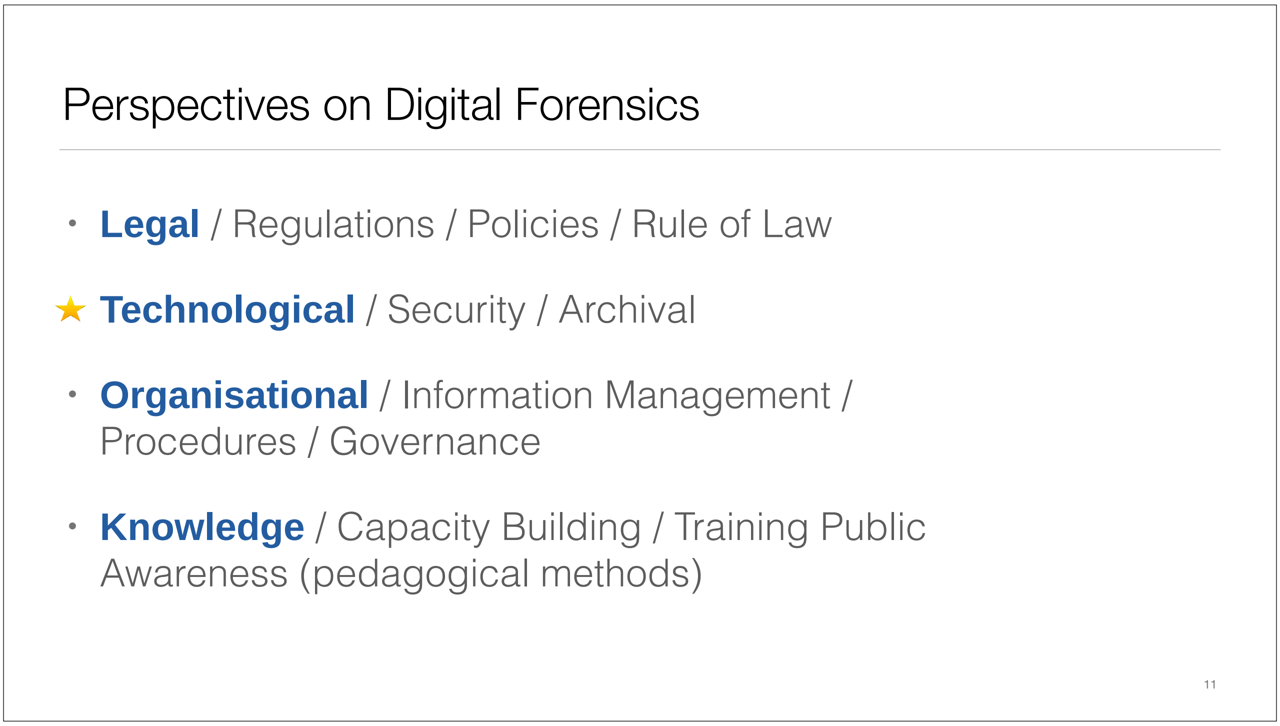
Nordic Computer Forensics Investigators Level 3 (NCFI 3)
(7.5 ETCS)

Experience-based Master in IS / Digital Forensics & Cybercrime Investigation (90 ETCS)

Master of Science in IS / Digital Forensics (120 ETCS)

PhD in IS / Digital Forensics (30 ETCS + Research)

10



Perspectives on Digital Forensics

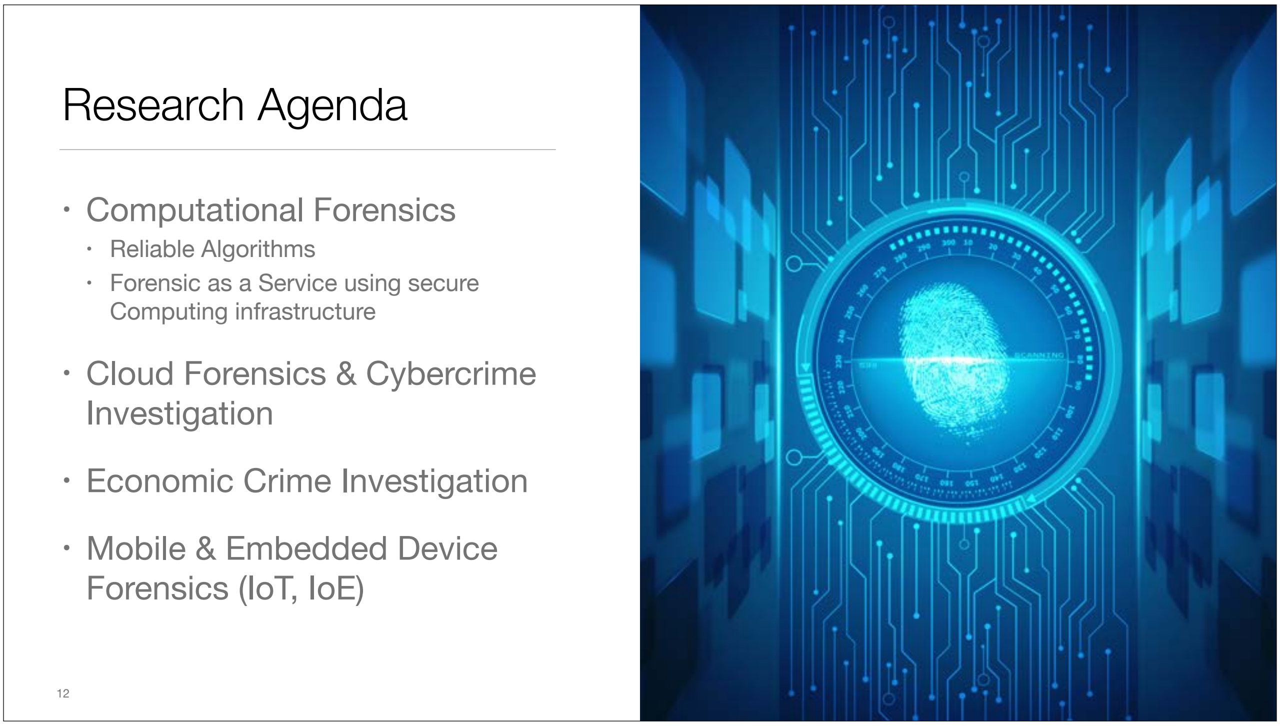
- **Legal** / Regulations / Policies / Rule of Law

★ **Technological** / Security / Archival

- **Organisational** / Information Management / Procedures / Governance
- **Knowledge** / Capacity Building / Training Public Awareness (pedagogical methods)

11

11



Research Agenda

- Computational Forensics
  - Reliable Algorithms
  - Forensic as a Service using secure Computing infrastructure
- Cloud Forensics & Cybercrime Investigation
- Economic Crime Investigation
- Mobile & Embedded Device Forensics (IoT, IoE)

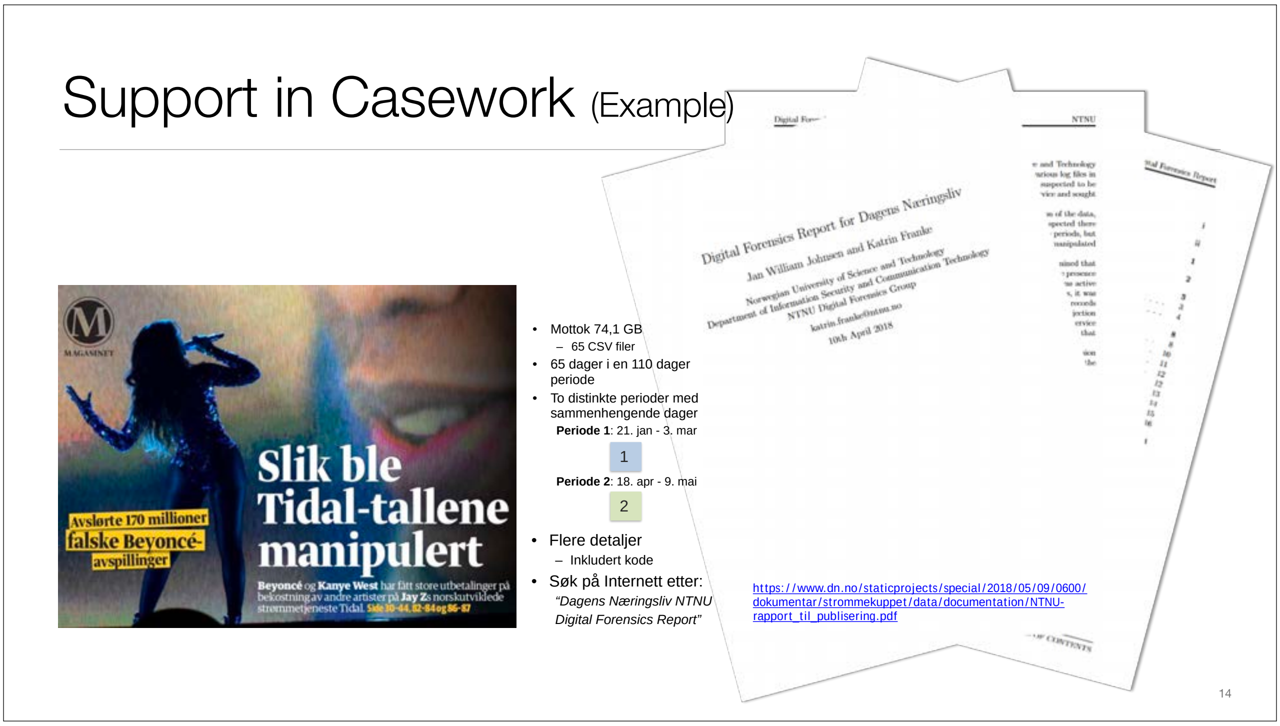
12

12

Impact & Return of Investments
(Examples)

- Additional External Funding, i.e. for 4 Police PhD
- Competence Increase within the Police
- Focused and in-depth Research and Innovation
- Access to International Academic Networks
- Assistance in High-tech Crime Casework

13



Support in Casework (Example)

M
MAGASINET
Avslørte 170 millioner falske Beyoncé-avspillinger
Slik ble Tidal-tallene manipulert

Beyoncé og **Kanye West** har fått store utbetalinger på bekostning av andre artister på **Jay Zs** norskutviklede strømmetjeneste Tidal. Side 30-44, 82-84 og 86-87

- Mottok 74,1 GB
  - 65 CSV filer
- 65 dager i en 110 dager periode
- To distinkte perioder med sammenhengende dager

  Periode 1: 21. jan - 3. mar

  1

  Periode 2: 18. apr - 9. mai

  2
- Flere detaljer
  - Inkludert kode
- Søk på Internett etter:

  "Dagens Næringsliv NTNU Digital Forensics Report"

Digital Forensics Report for Dagens Næringsliv

Jan William Johnsen and Katrin Franke

Norwegian University of Science and Technology
Department of Information Security and Communication Technology
NTNU Digital Forensics Group

katrin.franke@ntnu.no

10th April 2018

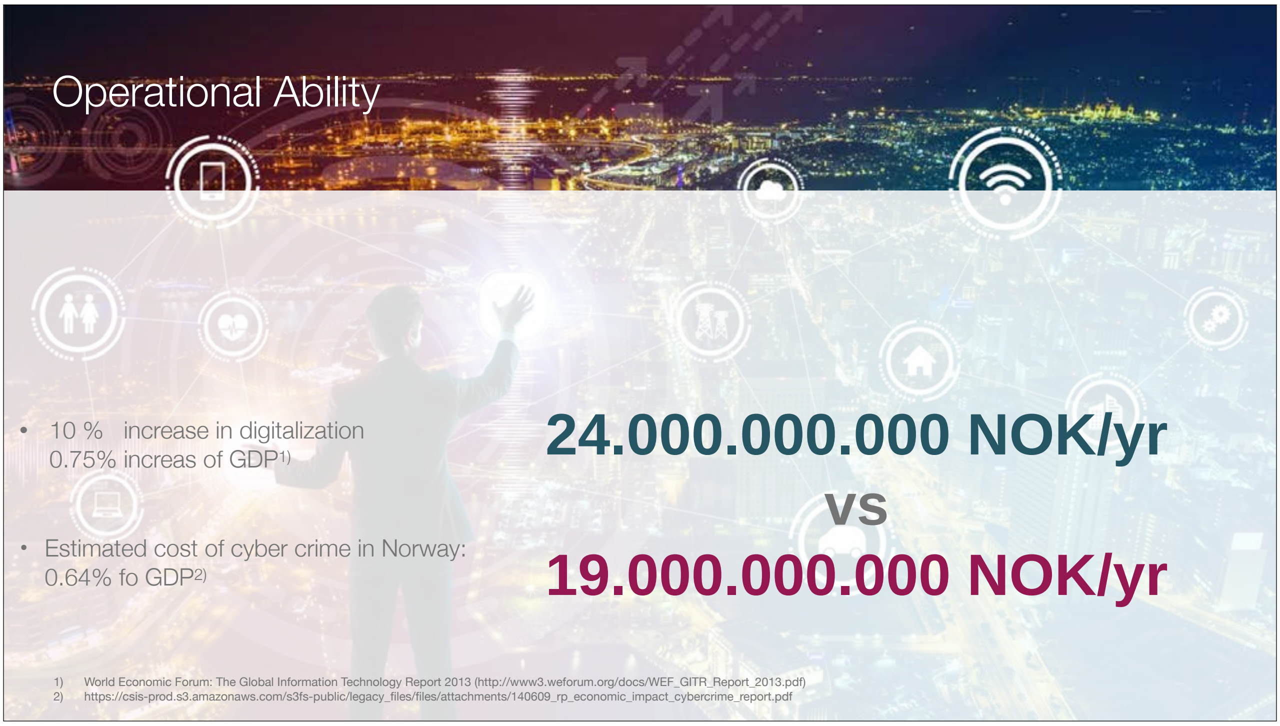
https://www.dn.no/staticprojects/special/2018/05/09/0600/dokumentar/strommekuppet/data/documentation/NTNU-rapport_til_publisering.pdf

14

14

This publication was realised with the EEA Financial Mechanism 2014-2021 financial support. Its content (text, photos, videos) does not reflect the official opinion of the Programme Operator, the National Contact Point and the Financial Mechanism Office. Responsibility for the information and views expressed therein lies entirely with the author(s).

Digitalisation of our Society



Operational Ability

- 10 % increase in digitalization
  0.75% increas of GDP1)
- Estimated cost of cyber crime in Norway:
  0.64% fo GDP2)

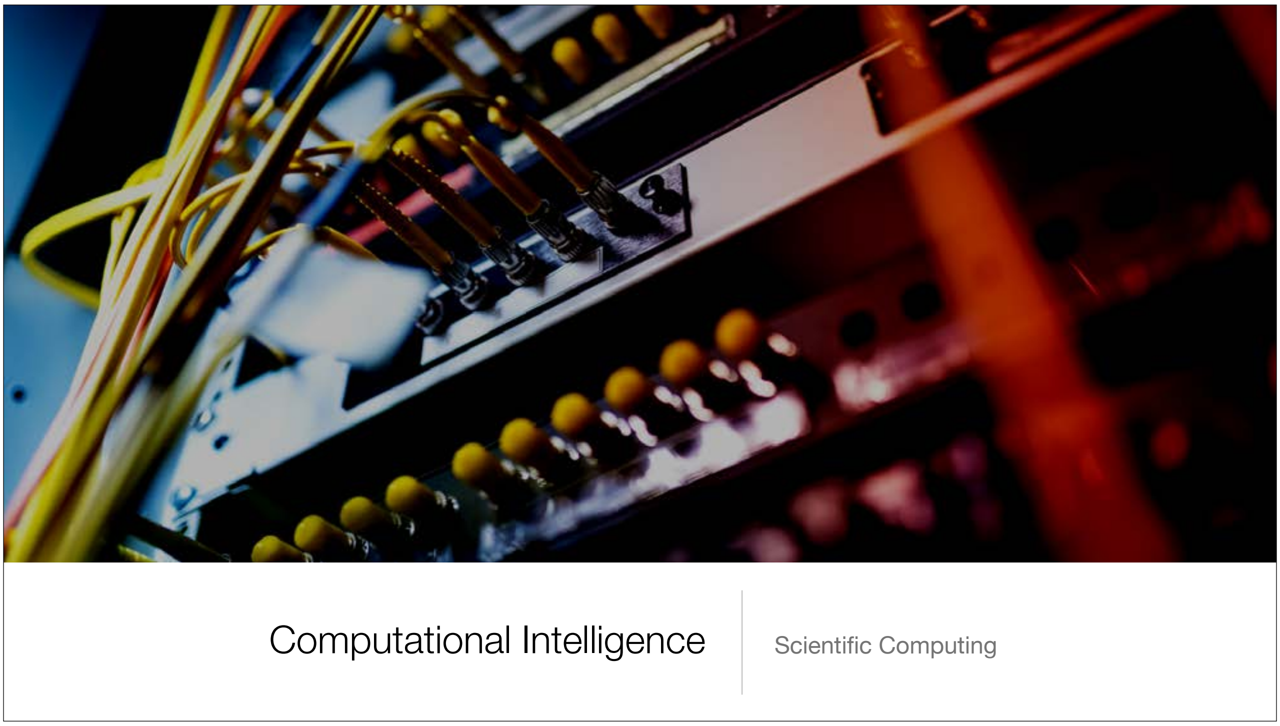
24.000.000.000 NOK/yr

vs

19.000.000.000 NOK/yr

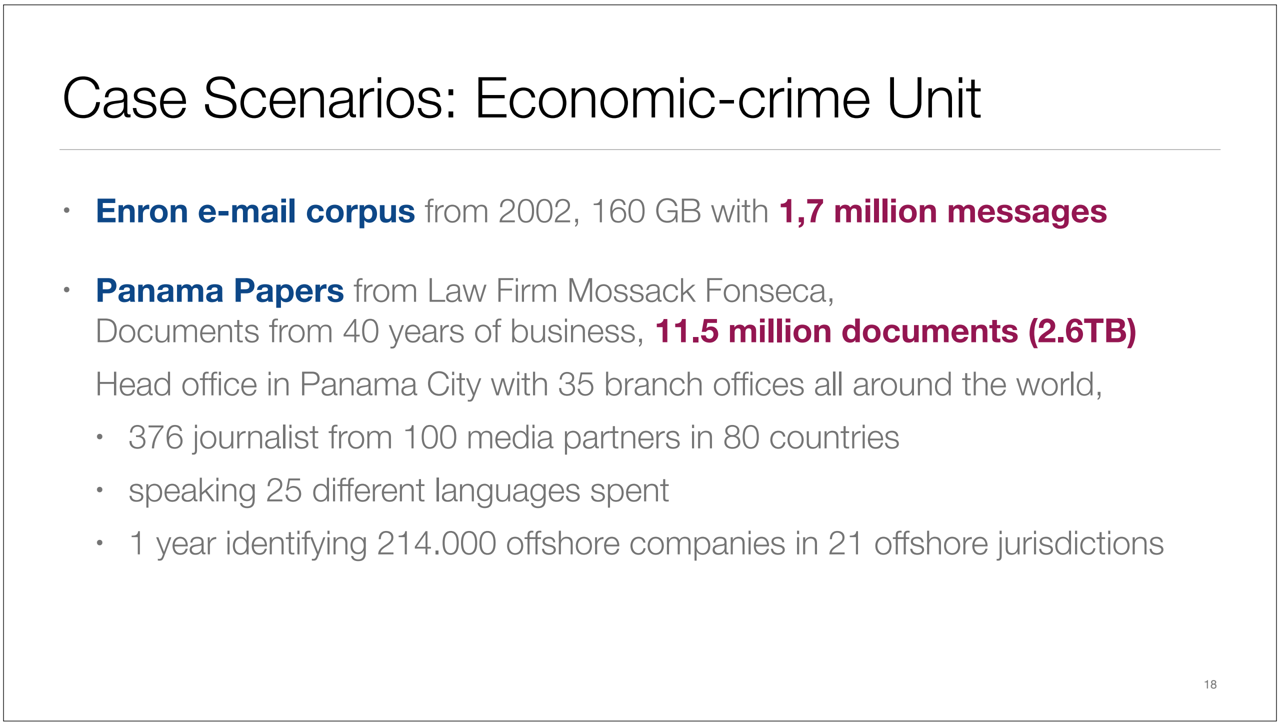
1) World Economic Forum: The Global Information Technology Report 2013 (http://www3.weforum.org/docs/WEF_GITR_Report_2013.pdf)
2) https://csis-prod.s3.amazonaws.com/s3fs-public/legacy_files/files/attachments/140609_rp_economic_impact_cybercrime_report.pdf

This publication was realised with the EEA Financial Mechanism 2014-2021 financial support. Its content (text, photos, videos) does not reflect the official opinion of the Programme Operator, the National Contact Point and the Financial Mechanism Office. Responsibility for the information and views expressed therein lies entirely with the author(s).



Computational Intelligence | Scientific Computing

17

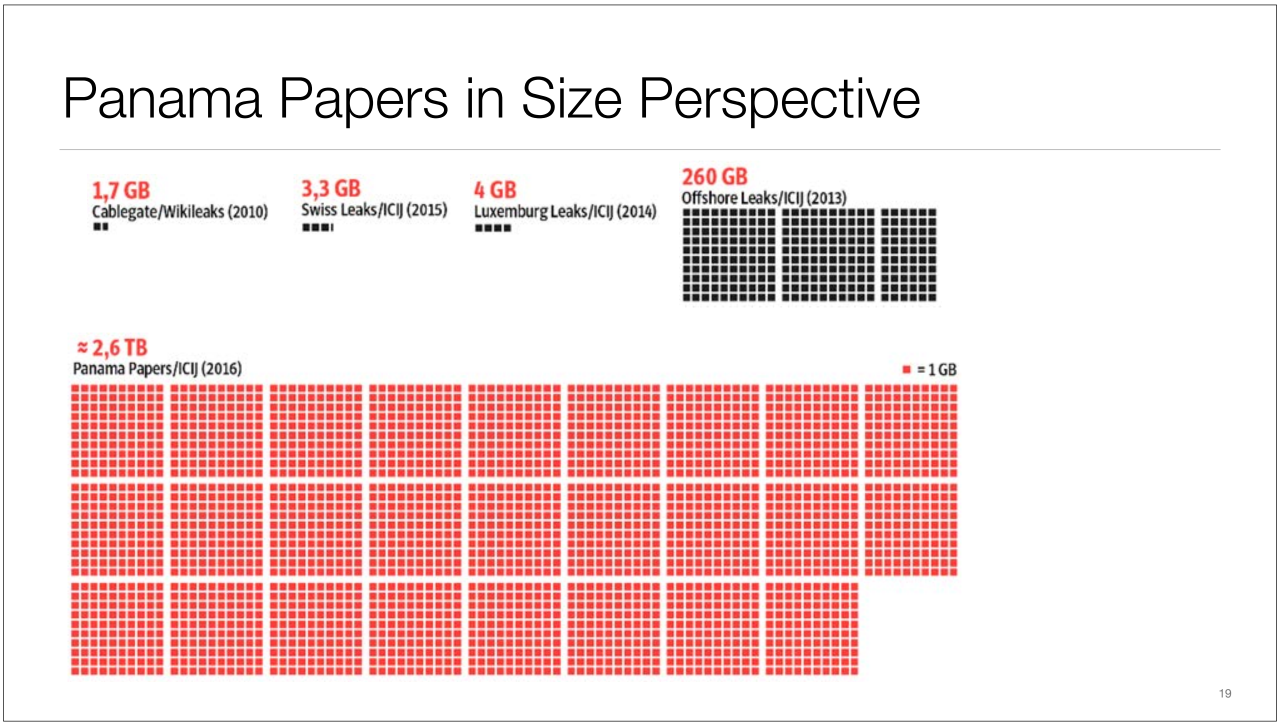


Case Scenarios: Economic-crime Unit

- **Enron e-mail corpus** from 2002, 160 GB with **1,7 million messages**
- **Panama Papers** from Law Firm Mossack Fonseca,
  Documents from 40 years of business, **11.5 million documents (2.6TB)**
  Head office in Panama City with 35 branch offices all around the world,
  - 376 journalist from 100 media partners in 80 countries
  - speaking 25 different languages spent
  - 1 year identifying 214.000 offshore companies in 21 offshore jurisdictions

18

18



Panama Papers in Size Perspective

1,7 GB Cablegate/Wikileaks (2010)

3,3 GB Swiss Leaks/ICIJ (2015)

4 GB Luxemburg Leaks/ICIJ (2014)

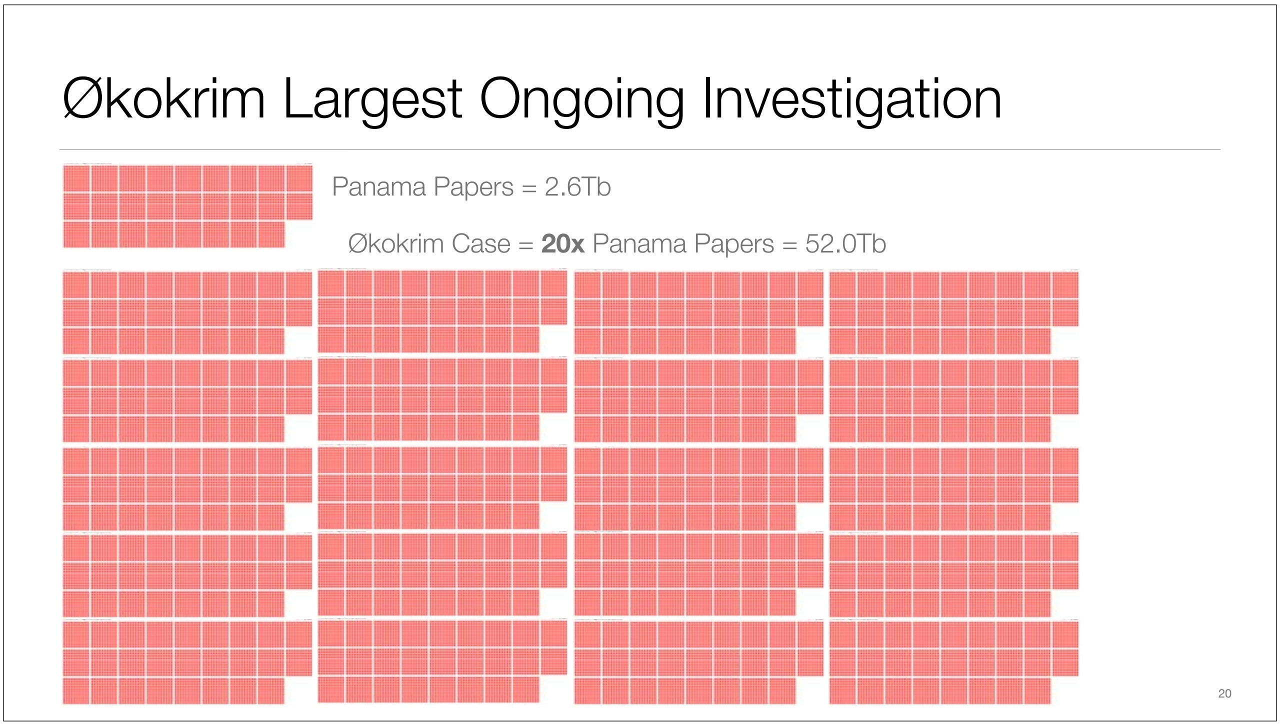
260 GB Offshore Leaks/ICIJ (2013)

≈ 2,6 TB Panama Papers/ICIJ (2016)

■ = 1 GB

19

19



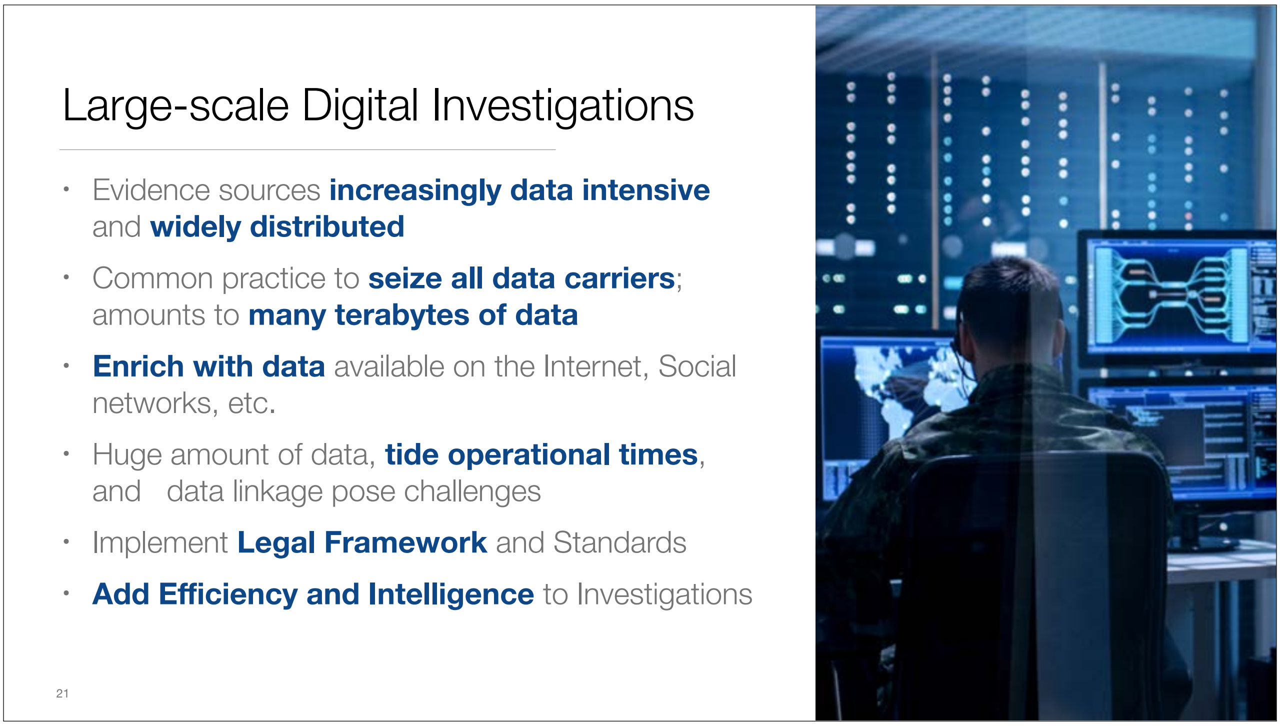
Økokrim Largest Ongoing Investigation

Panama Papers = 2.6Tb

Økokrim Case = **20x** Panama Papers = 52.0Tb

20

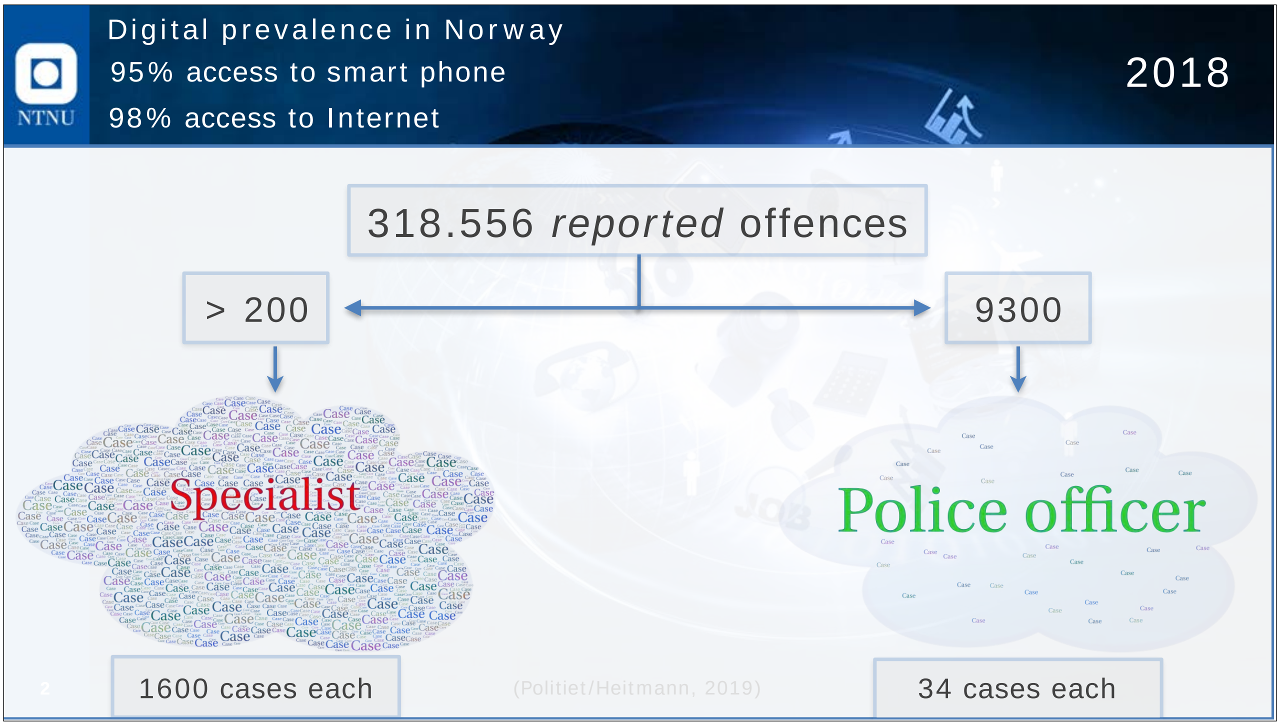
20



Large-scale Digital Investigations

- Evidence sources **increasingly data intensive** and **widely distributed**
- Common practice to **seize all data carriers**; amounts to **many terabytes of data**
- **Enrich with data** available on the Internet, Social networks, etc.
- Huge amount of data, **tide operational times**, and data linkage pose challenges
- Implement **Legal Framework** and Standards
- **Add Efficiency and Intelligence** to Investigations

21



Digital prevalence in Norway
95% access to smart phone
98% access to Internet

2018

318.556 *reported* offences

> 200

9300

Specialist

Police officer

1600 cases each

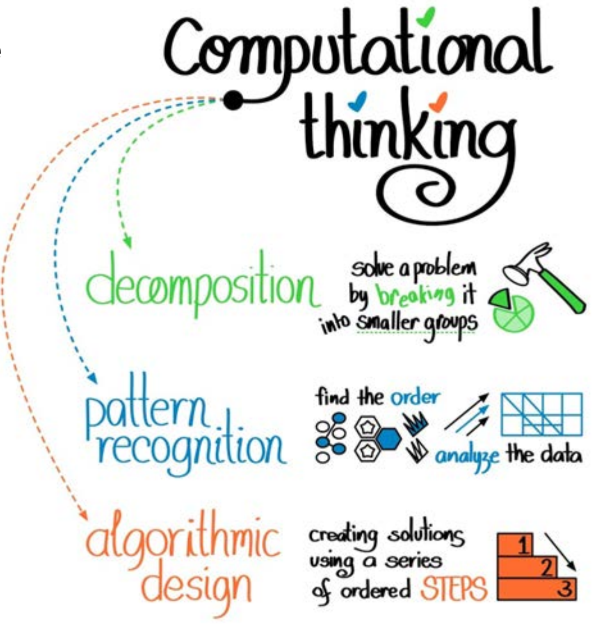
(Politiet/Heitmann, 2019)

34 cases each

This publication was realised with the EEA Financial Mechanism 2014-2021 financial support. Its content (text, photos, videos) does not reflect the official opinion of the Programme Operator, the National Contact Point and the Financial Mechanism Office. Responsibility for the information and views expressed therein lies entirely with the author(s).

A digital age skill for everyone

- https://www.youtube.com/watch?v=VFcUgSYyRPg
- https://www.youtube.com/watch?v=mUXo-S7gzds
- https://www.youtube.com/watch?v=AkzdvKhbWLQ



PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A

Phil. Trans. R. Soc. A (2008) 366, 3717–3725
doi:10.1098/rsta.2008.0118
Published online 31 July 2008

Computational thinking and thinking about computing

By Jeannette M. Wing*

Computer Science Department, Carnegie Mellon University, Pittsburgh, PA 15213, USA

Computational thinking will influence everyone in every field of endeavour. This vision poses a new educational challenge for our society, especially for our children. In thinking about computing, we need to be attuned to the three drivers of our field: science, technology and society. Accelerating technological advances and monumental societal demands force us to revisit the most basic scientific questions of computing.

Keywords: computational thinking; abstraction; automation; computing; computable; intelligence

1. Computational thinking

Computational thinking is taking an approach to solving problems, designing systems and understanding human behaviour that draws on concepts fundamental to computing[1] (Wing 2006).

Computational thinking is a kind of analytical thinking. It shares with mathematical thinking in the general ways in which we might approach solving a problem. It shares with engineering thinking in the general ways in which we might approach designing and evaluating a large, complex system that operates within the constraints of the real world. It shares with scientific thinking in the general ways in which we might approach understanding computability, intelligence, the mind and human behaviour.

(a) Computing: abstraction and automation

The essence of computational thinking is abstraction. In computing, we abstract notions beyond the physical dimensions of time and space. Our abstractions are extremely general because they are symbolic, where numeric abstractions are just a special case.

In two ways, our abstractions tend to be richer and more complex than those in the mathematical and physical sciences. First, our abstractions do not necessarily enjoy the clean, elegant or easily definable algebraic properties of mathematical

*wing@cs.cmu.edu

[1] By 'computing' I mean very broadly the field encompassing computer science, computer engineering, communications, information science and information technology.

One contribution of 19 to a Discussion Meeting Issue 'From computers to ubiquitous computing, by 2020'.

3717 This journal is © 2008 The Royal Society

Computational Forensics: An Overview

Katrin Franke[1] and Sargur N. Srihari[2]

[1] Norwegian Information Security Laboratory, Gjøvik University College, Norway
[2] CEDAR, University at Buffalo, State University of New York, USA
kyfranke@ieee.org,
srihari@cedar.buffalo.edu

Abstract. Cognitive abilities of human expertise modeled using computational methods offer several new possibilities for the forensic sciences. They include three areas: providing tools for use by the forensic examiner, establishing a scientific basis for the expertise, and providing an alternate opinion on a case. This paper gives a brief overview of computational forensics with a focus on those disciplines that involve pattern evidence.

Keywords: Computational science, Forensic science, Computer science, Artificial intelligence, Law enforcement, Investigation services.

1 Introduction

The term "computational" has been associated with several disciplines of human expertise. Examples are computational vision, computational linguistics, computational chemistry, computational advertising, etc. Analogously a body of knowledge and methods to be collectively defined as computational forensics can be defined.

Computational methods find a place in the forensic sciences in three ways. First, they provide tools for the human examiner to better analyze evidence by overcoming limitations of human cognitive ability– thus they can support the forensic examiner in his / her daily casework. Secondly they can be used to provide the scientific basis for a forensic discipline or procedure by providing for the analysis of large volumes of data which are not humanly possible. Thirdly they can ultimately be used to represent human expert knowledge and for implementing recognition and reasoning abilities in machines. While the goal of a computer to provide an opinion is a goal analogous to other grand challenges of artificial intelligence, they are unlikely to replace the human examiner in the foreseeable future. On the other hand it is more likely that modern crime investigation will profit from the hybrid-intelligence of humans and machines.

More broadly, computer methods and algorithms enable the forensic practitioner to:

– reveal and improve traces evidence for further investigation,
– analyze and identify evidence in an objective and reproducible manner,

S.N. Srihari and K. Franke (Eds.): IWCF 2008, LNCS 5158, pp. 1–10, 2008.
© Springer-Verlag Berlin Heidelberg 2008

STRENGTHENING
FORENSIC SCIENCE
IN THE UNITED STATES

A PATH FORWARD

Committee on Identifying the Needs of the Forensic Science Community

Committee on Science, Technology, and Law
Policy and Global Affairs

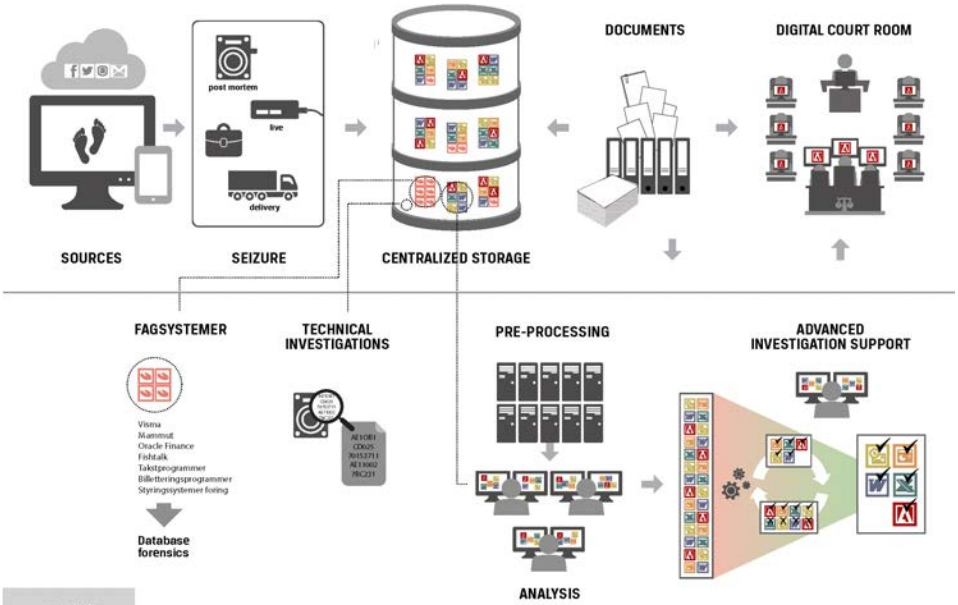
Committee on Applied and Theoretical Statistics
Division on Engineering and Physical Sciences

NATIONAL RESEARCH COUNCIL
OF THE NATIONAL ACADEMIES

Computational Methods

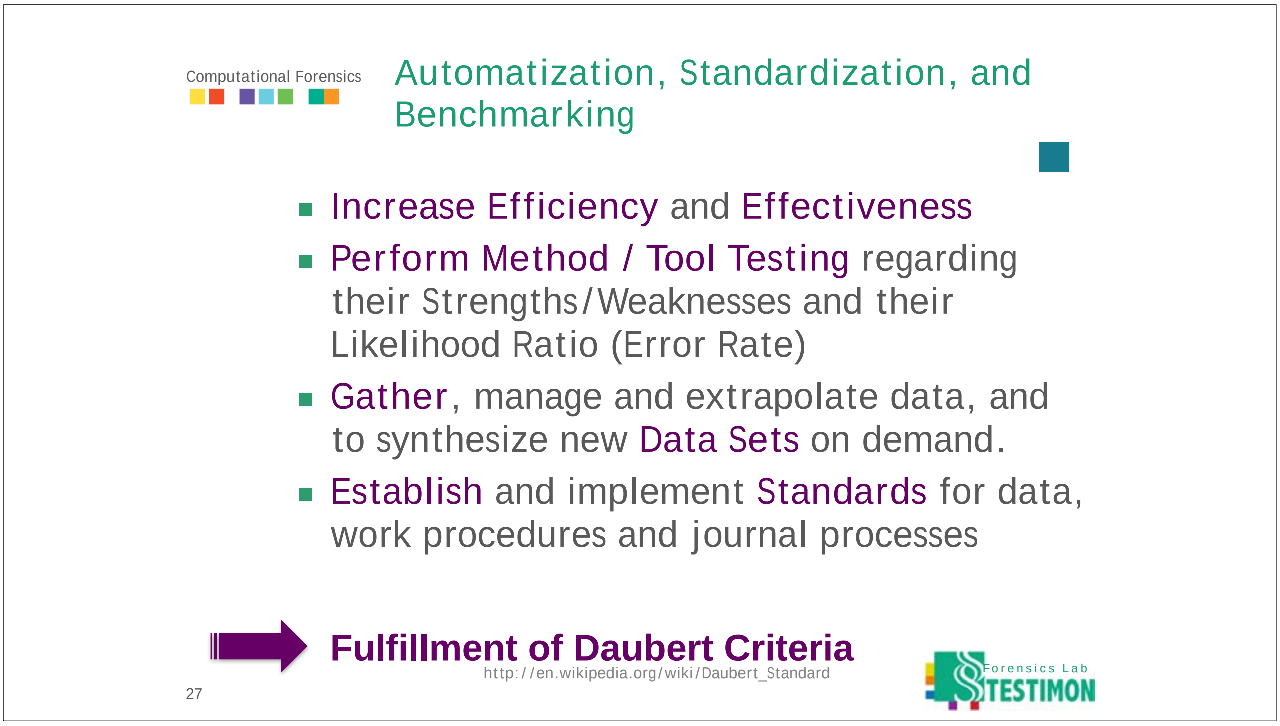
- **Signal / Image Processing** : one-dimensional signals and two-dimensional images are transformed for better human or machine processing,
- **Computer Vision** : images are automatically recognised to identify objects,
- **Computer Graphics / Data Visualisation** : two-dimensional images or three-dimensional scenes are synthesised from multi-dimensional data for better human understanding,
- **Statistical Pattern Recognition** : abstract measurements are classified as belonging to one or more classes, e.g., whether a sample belongs to a known class and with what probability,
- **Machine Learning** : a mathematical model is learnt from examples.
- **Data Mining** : large volumes of data are processed to discover nuggets of information, e.g., presence of associations, number of clusters, outliers, etc.
- **Robotics** : human movements are replicated by a machine.

Project Example - Ars Forensica



Designed by/for ØKOKRIM NTNU CCIS

This publication was realised with the EEA Financial Mechanism 2014-2021 financial support. Its content (text, photos, videos) does not reflect the official opinion of the Programme Operator, the National Contact Point and the Financial Mechanism Office. Responsibility for the information and views expressed therein lies entirely with the author(s).



Computational Forensics

Automatization, Standardization, and Benchmarking

- Increase Efficiency and Effectiveness
- Perform Method / Tool Testing regarding their Strengths/Weaknesses and their Likelihood Ratio (Error Rate)
- Gather, manage and extrapolate data, and to synthesize new Data Sets on demand.
- Establish and implement Standards for data, work procedures and journal processes

➡ **Fulfillment of Daubert Criteria**

http://en.wikipedia.org/wiki/Daubert_Standard

27

Code-breaking Enigma, December 1942

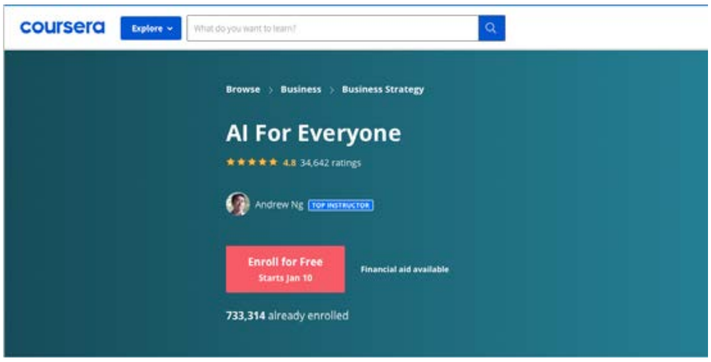
This publication was realised with the EEA Financial Mechanism 2014-2021 financial support. Its content (text, photos, videos) does not reflect the official opinion of the Programme Operator, the National Contact Point and the Financial Mechanism Office. Responsibility for the information and views expressed therein lies entirely with the author(s).

Computing Machines & Intelligence (1950) by

Alan Turing

https://wsimag.com/science-and-technology/36961-no-turing-test-for-consciousness

Artificial Intelligence for Everyone



Browse › Business › Business Strategy

AI For Everyone

★★★★★ 4.8 34,642 ratings

Andrew Ng TOP INSTRUCTOR

Enroll for Free
Starts Jan 10

Financial aid available

733,314 already enrolled

About this Course

AI is not only for engineers. If you want your organisation to become better at using AI, this is the course to tell everyone--especially your non-technical colleagues--to take.

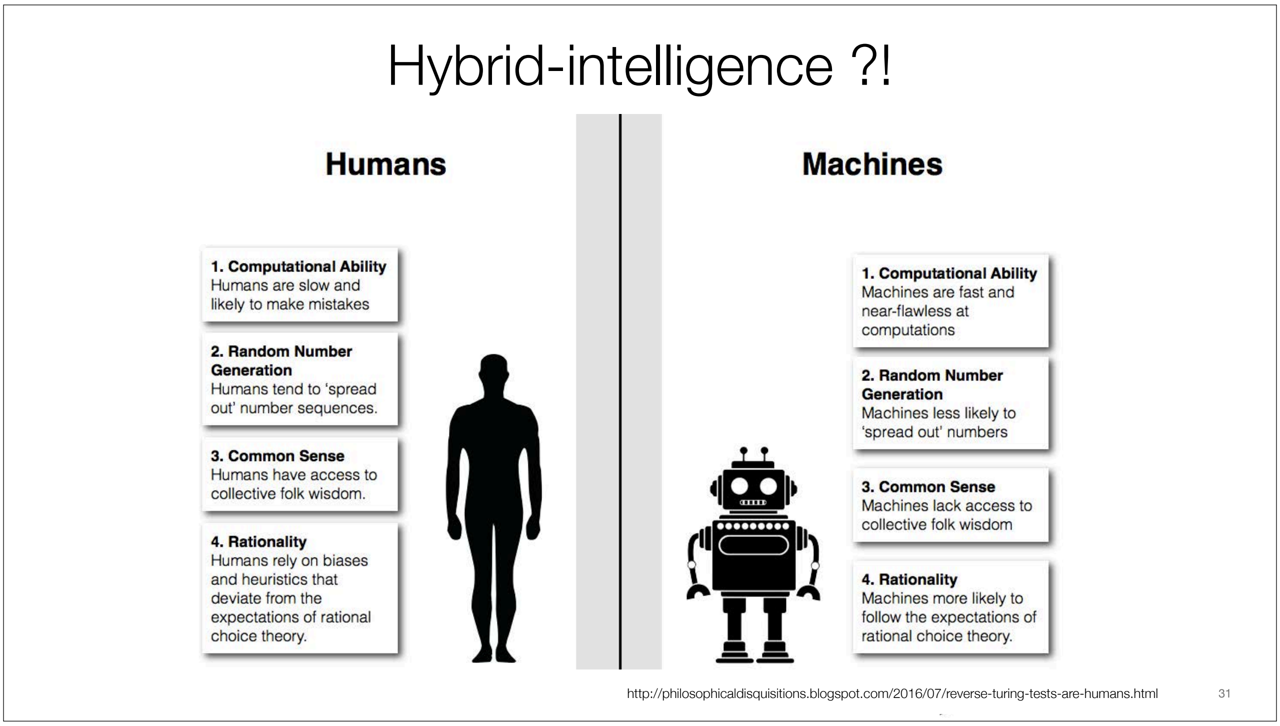
In this course, you will learn:

- The meaning behind common AI terminology, including neural networks, machine learning, deep learning, and data science
- What AI realistically can--and cannot--do
- How to spot opportunities to apply AI to problems in your own organisation
- What it feels like to build machine learning and data science projects
- How to work with an AI team and build an AI strategy in your company
- How to navigate ethical and societal discussions surrounding AI

Though this course is largely non-technical, engineers can also take this course to learn the business aspects of AI.

Link: https://tinyurl.com/AI-4-Everyone

This publication was realised with the EEA Financial Mechanism 2014-2021 financial support. Its content (text, photos, videos) does not reflect the official opinion of the Programme Operator, the National Contact Point and the Financial Mechanism Office. Responsibility for the information and views expressed therein lies entirely with the author(s).



Hybrid-intelligence ?!

Humans

1. Computational Ability
Humans are slow and likely to make mistakes

2. Random Number Generation
Humans tend to 'spread out' number sequences.

3. Common Sense
Humans have access to collective folk wisdom.

4. Rationality
Humans rely on biases and heuristics that deviate from the expectations of rational choice theory.

Machines

1. Computational Ability
Machines are fast and near-flawless at computations

2. Random Number Generation
Machines less likely to 'spread out' numbers

3. Common Sense
Machines lack access to collective folk wisdom

4. Rationality
Machines more likely to follow the expectations of rational choice theory.

http://philosophicaldisquisitions.blogspot.com/2016/07/reverse-turing-tests-are-humans.html 31

31

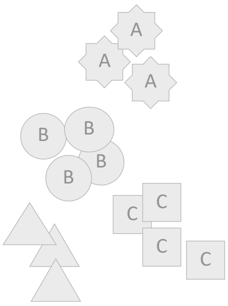
Machine Learning & Pattern Recognition

Fundamentals

32

This publication was realised with the EEA Financial Mechanism 2014-2021 financial support. Its content (text, photos, videos) does not reflect the official opinion of the Programme Operator, the National Contact Point and the Financial Mechanism Office. Responsibility for the information and views expressed therein lies entirely with the author(s).

Machine Learning & Pattern Recognition



Pattern

- "as opposite of a chaos;
  it is an entity, vaguely defined, that
  could be given a name" Watanabe 1985

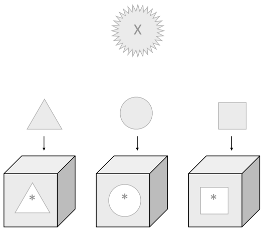
Goals

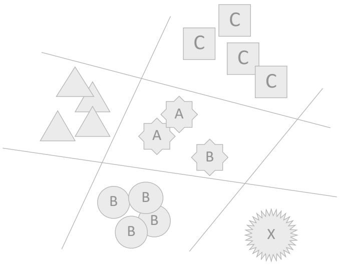
- Supervised / Unsupervised Classification of Patterns by means of Computational Methods
- Small Intra-class & Large Inter-class Variation

Same Facet - Different Origin

- Machine Learning - Computer Science
- Patter Recognition / Data Mining - Engineering
- Predictive Analytics - Business / Marketing

Pattern Classification





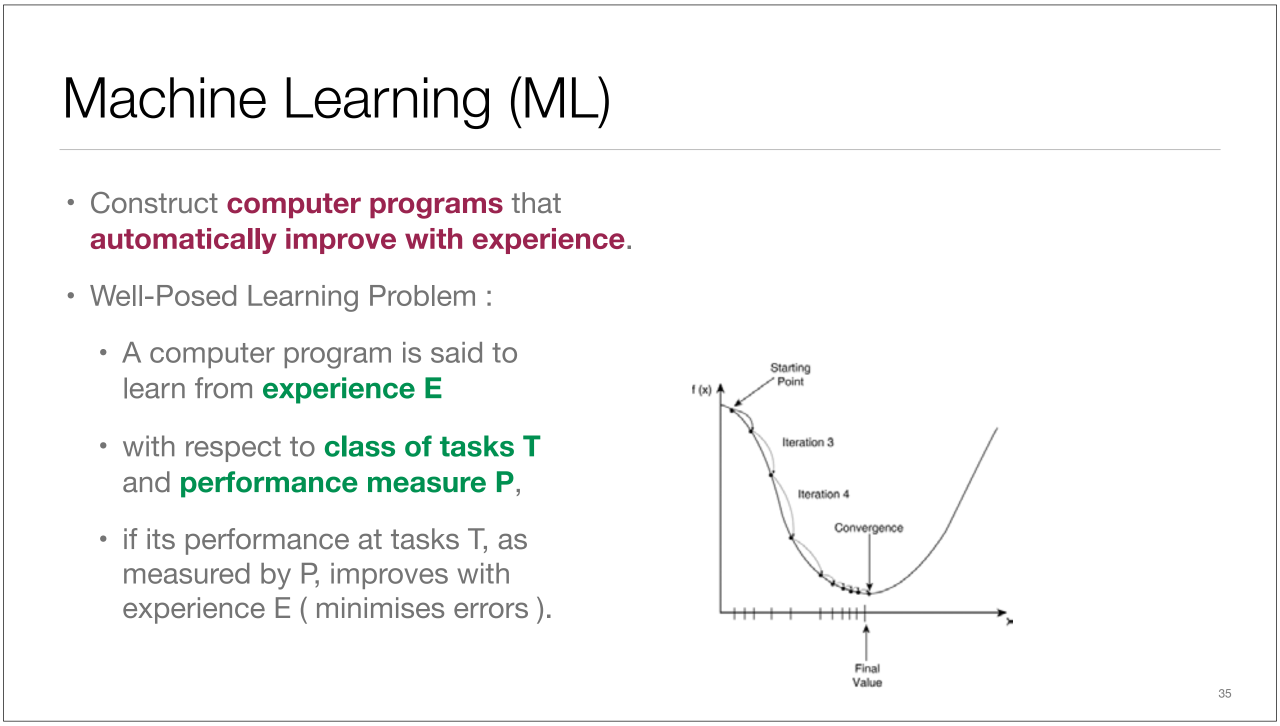
Supervised Classification pre-defined by the system designer

Machine Learning

Unsupervised Classification learning based on the similarity of pattern

Data Mining

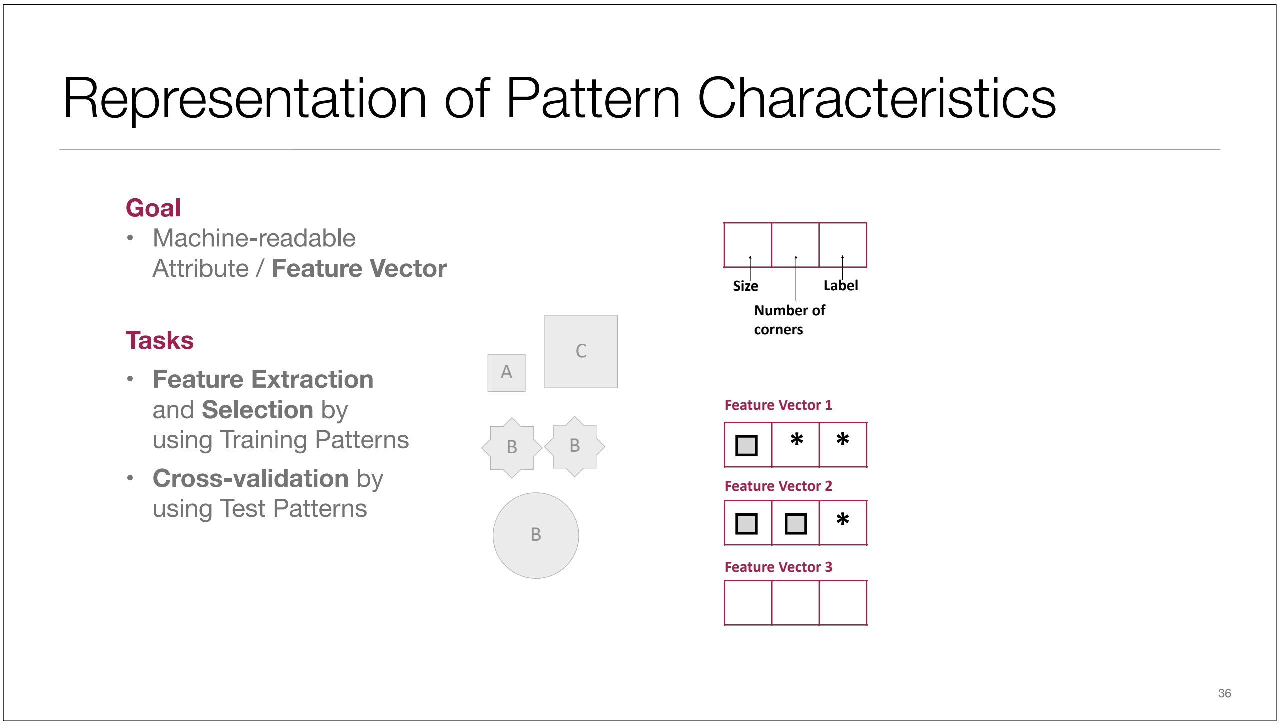
This publication was realised with the EEA Financial Mechanism 2014-2021 financial support. Its content (text, photos, videos) does not reflect the official opinion of the Programme Operator, the National Contact Point and the Financial Mechanism Office. Responsibility for the information and views expressed therein lies entirely with the author(s).



Machine Learning (ML)

- Construct **computer programs** that **automatically improve with experience**.
- Well-Posed Learning Problem :
  - A computer program is said to learn from **experience E**
  - with respect to **class of tasks T** and **performance measure P**,
  - if its performance at tasks T, as measured by P, improves with experience E (minimises errors).

f (x)
Starting Point
Iteration 3
Iteration 4
Convergence
Final Value



Representation of Pattern Characteristics

Goal

- Machine-readable Attribute / **Feature Vector**

Tasks

- **Feature Extraction** and **Selection** by using Training Patterns
- **Cross-validation** by using Test Patterns

C
A
B
B
B
Size
Number of corners
Label

Feature Vector 1

□	*	*

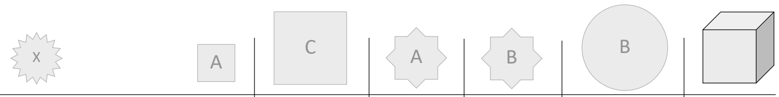
Feature Vector 2

□	□	*

Feature Vector 3

This publication was realised with the EEA Financial Mechanism 2014-2021 financial support. Its content (text, photos, videos) does not reflect the official opinion of the Programme Operator, the National Contact Point and the Financial Mechanism Office. Responsibility for the information and views expressed therein lies entirely with the author(s).

Pattern Representation & Classification



	X	A	C	A	B	B	
Feature Vector 1 □ * *		1**	2**	1**	1**	2**	2 (2)
Feature Vector 2 □ □ *		14*	24*	16*	16*	20*	4 (6)
Feature Vector 3 □ □ □		14A	24C	16A	16B	20B	5 (18)

Size — Number of corners — Label

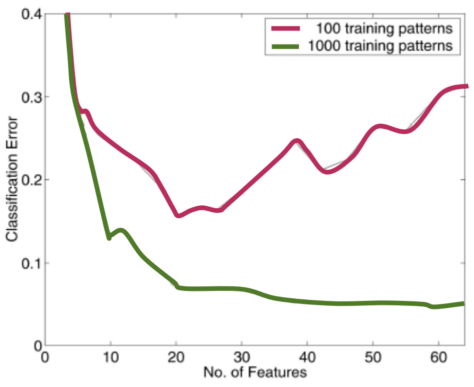
Classes

37

37

Classifier Training, … How do Computers learn?

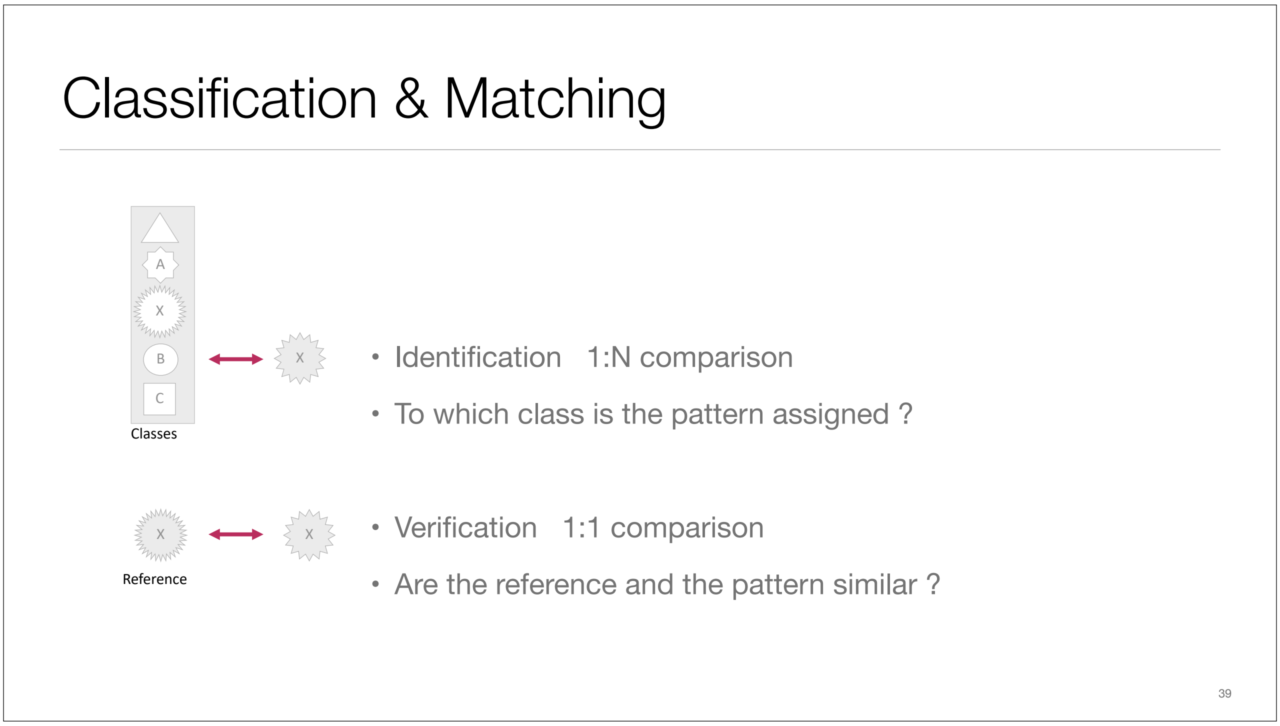
- Learning by Example !
- Requirements
  - Representative Sample Data
  - Appropriate Feature Encoding
- Challenge
  - Class Discrimination
  - Avoid Over Learning



38

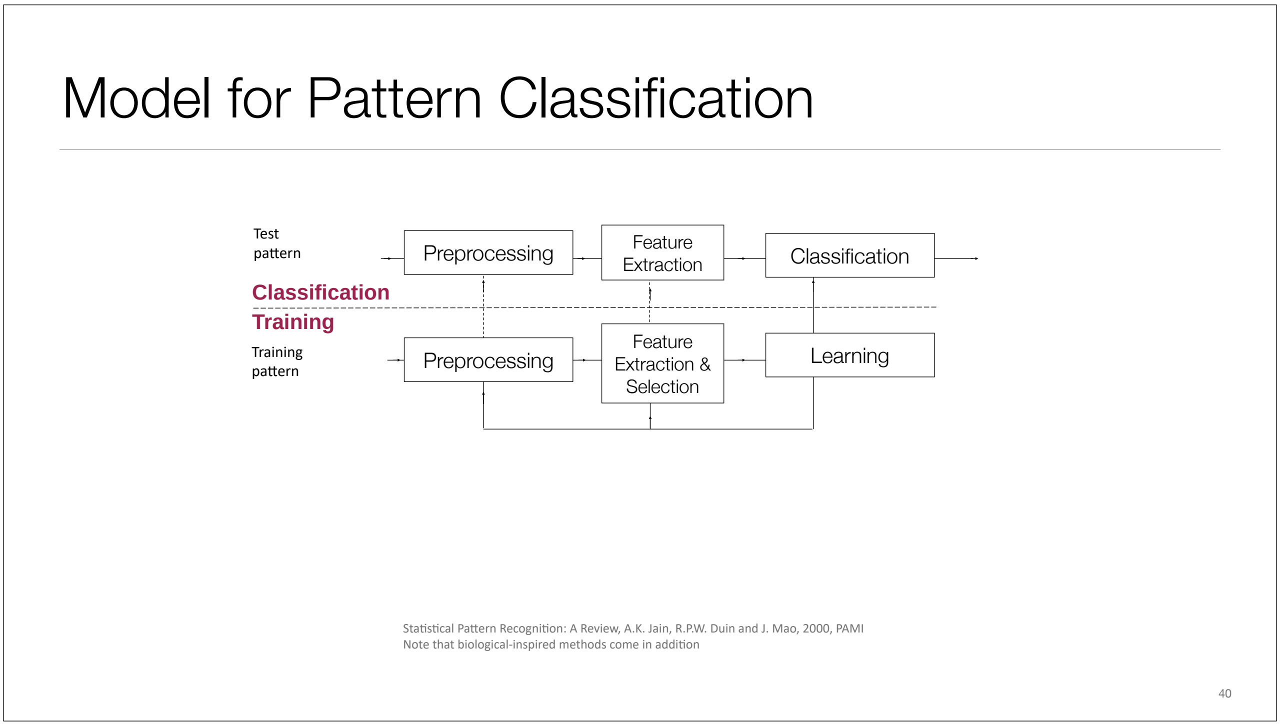
38

This publication was realised with the EEA Financial Mechanism 2014-2021 financial support. Its content (text, photos, videos) does not reflect the official opinion of the Programme Operator, the National Contact Point and the Financial Mechanism Office. Responsibility for the information and views expressed therein lies entirely with the author(s).



Classification & Matching

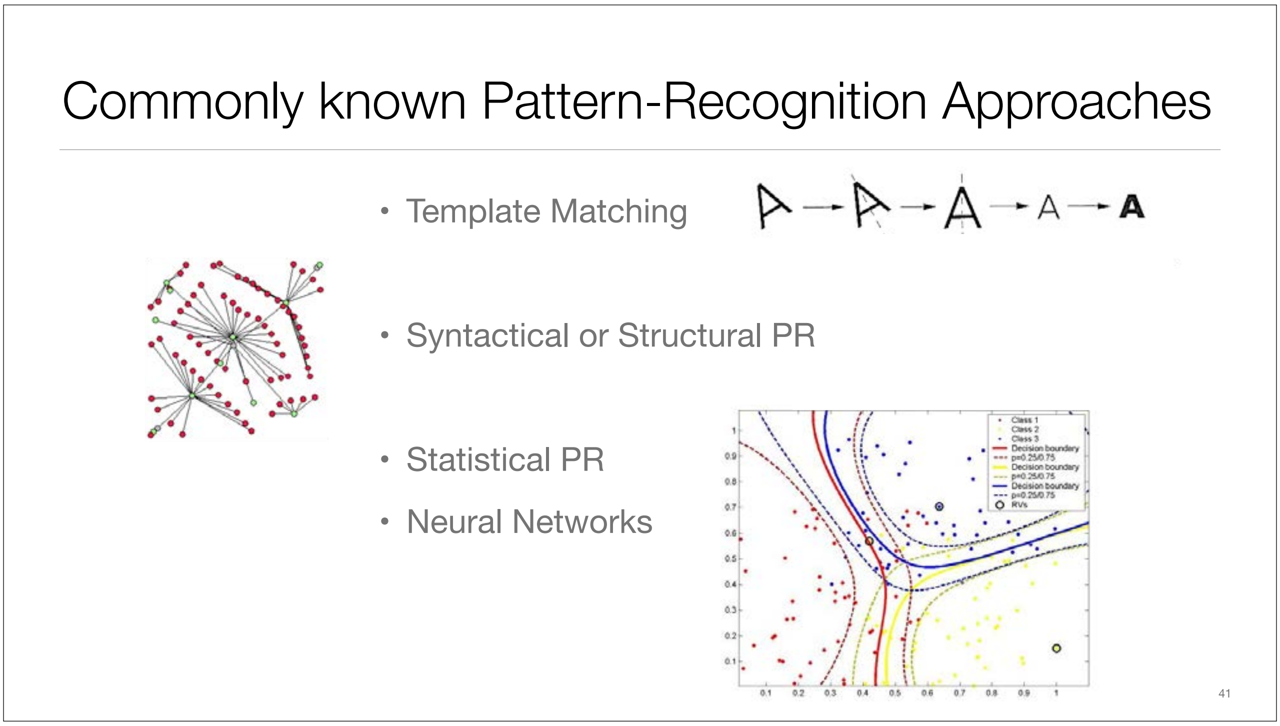
- Identification   1:N comparison
- To which class is the pattern assigned ?
- Verification   1:1 comparison
- Are the reference and the pattern similar ?



Model for Pattern Classification

Statistical Pattern Recognition: A Review, A.K. Jain, R.P.W. Duin and J. Mao, 2000, PAMI
Note that biological-inspired methods come in addition

This publication was realised with the EEA Financial Mechanism 2014-2021 financial support. Its content (text, photos, videos) does not reflect the official opinion of the Programme Operator, the National Contact Point and the Financial Mechanism Office. Responsibility for the information and views expressed therein lies entirely with the author(s).

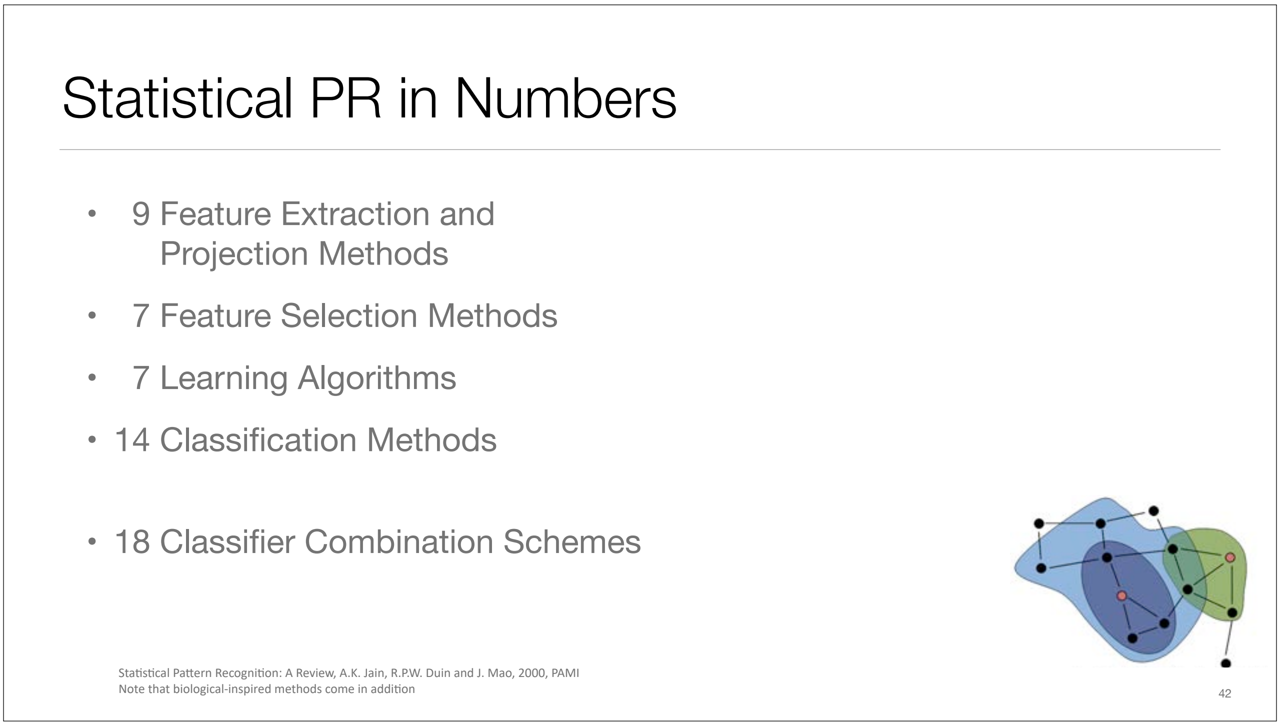


Commonly known Pattern-Recognition Approaches

- Template Matching
- Syntactical or Structural PR
- Statistical PR
- Neural Networks

Class 1
Class 2
Class 3
Decision boundary
p=0.25/0.75
Decision boundary
p=0.25/0.75
Decision boundary
p=0.25/0.75
RVs

41



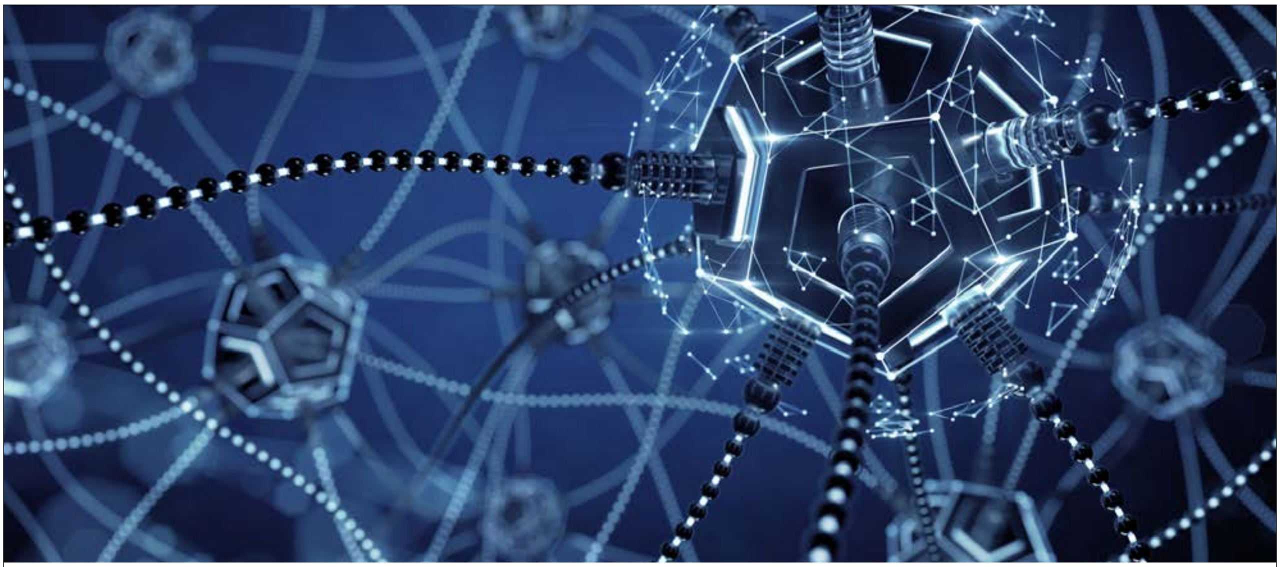
Statistical PR in Numbers

- 9 Feature Extraction and Projection Methods
- 7 Feature Selection Methods
- 7 Learning Algorithms
- 14 Classification Methods
- 18 Classifier Combination Schemes

Statistical Pattern Recognition: A Review, A.K. Jain, R.P.W. Duin and J. Mao, 2000, PAMI
Note that biological-inspired methods come in addition

42

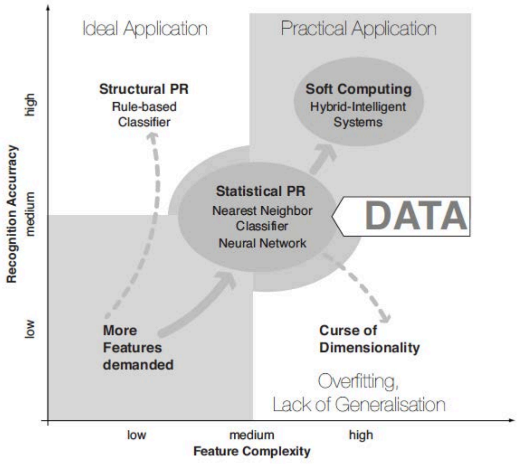
This publication was realised with the EEA Financial Mechanism 2014-2021 financial support. Its content (text, photos, videos) does not reflect the official opinion of the Programme Operator, the National Contact Point and the Financial Mechanism Office. Responsibility for the information and views expressed therein lies entirely with the author(s).



Data Science | Machine Learning & Computational Intelligence

43

Towards Data-driven Approaches



BIG DATA Analytics

Inter-relation of
feature complexity and
expected recognition accuracy.

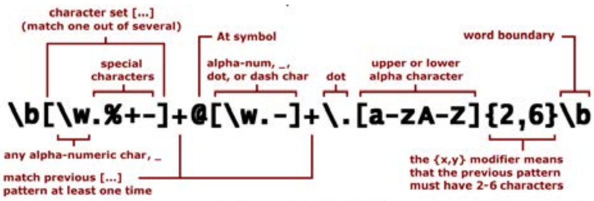
Reference: Franke (2005)

44

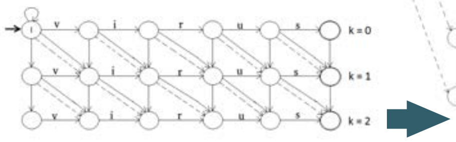
44

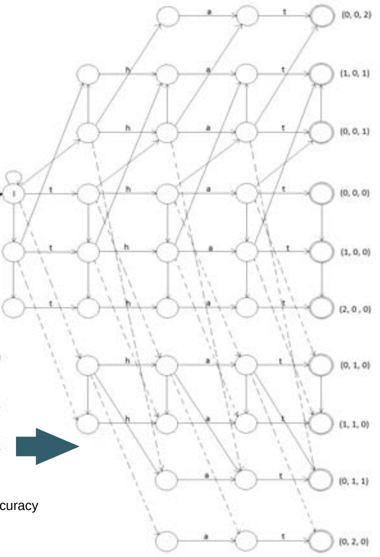
This publication was realised with the EEA Financial Mechanism 2014-2021 financial support. Its content (text, photos, videos) does not reflect the official opinion of the Programme Operator, the National Contact Point and the Financial Mechanism Office. Responsibility for the information and views expressed therein lies entirely with the author(s).

Regular Expressions vs. Approximate String Matching


character set [...]
(match one out of several)
special characters
At symbol
alpha-num, _, dot, or dash char
dot
upper or lower alpha character
word boundary
\b[\w.%+-]+@[\w.-]+\.[a-zA-Z]{2,6}\b
any alpha-numeric char, _
match previous [...] pattern at least one time
the {x,y} modifier means that the previous pattern must have 2-6 characters

Parse: username@domain.TLD (top level domain)



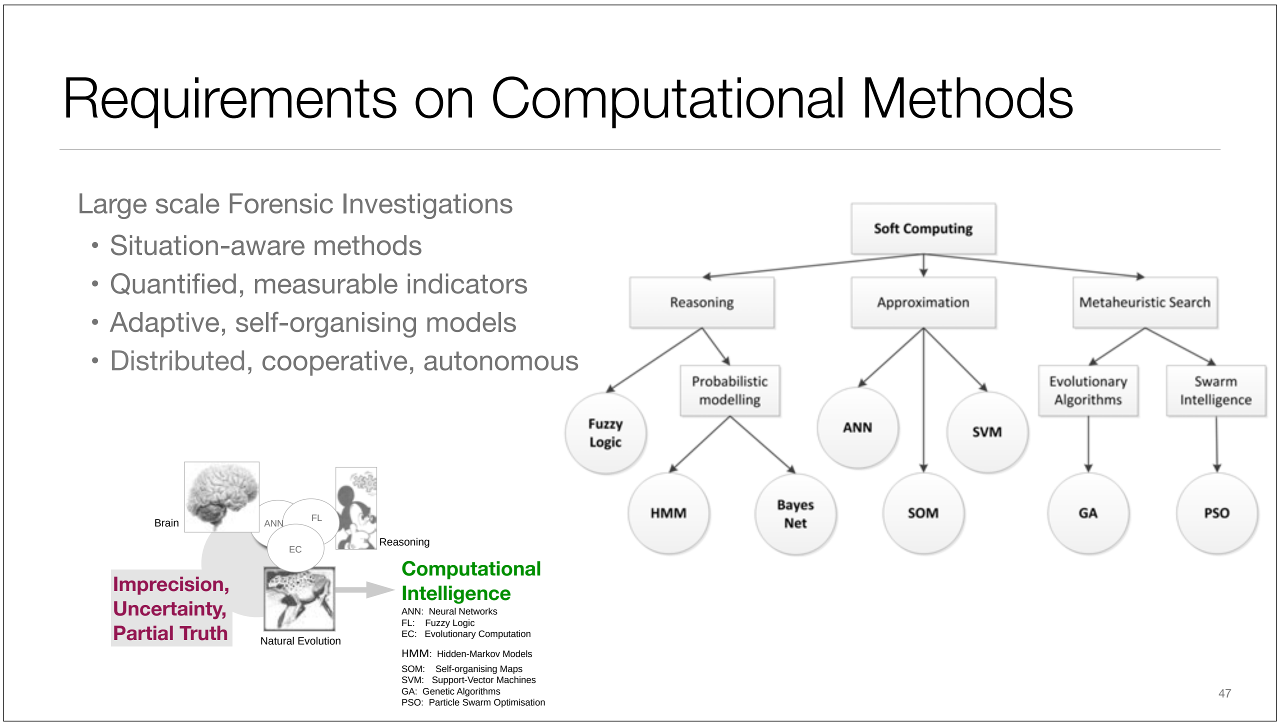


Improve precision in approximate (fuzzy) search
- Find more of what we want, without losing significant accuracy
- Good for beginning of investigation (K.Porter, 2017-)

Theoretical Foundations

- Algorithm Independent Means *(selection)*
  - **Ugly-Duckling Theorem**, S. Watanabe, 1969
    - Lack of any one feature or pattern representation that yields better classification performance without prior assumption
    - All differences are equal, unless one has some prior knowledge
  - **No-Free Lunch Theorem**, D.H. Wolpert and W.G. Macready, 1997
    - Lack of inherent superiority of any classifier
    - Q.: Which algorithm is suitable for which problem?
    - A.: Given an algorithm with an intended operating range R, it will be possible to find a problem in R which can not be be solved.

This publication was realised with the EEA Financial Mechanism 2014-2021 financial support. Its content (text, photos, videos) does not reflect the official opinion of the Programme Operator, the National Contact Point and the Financial Mechanism Office. Responsibility for the information and views expressed therein lies entirely with the author(s).



Requirements on Computational Methods

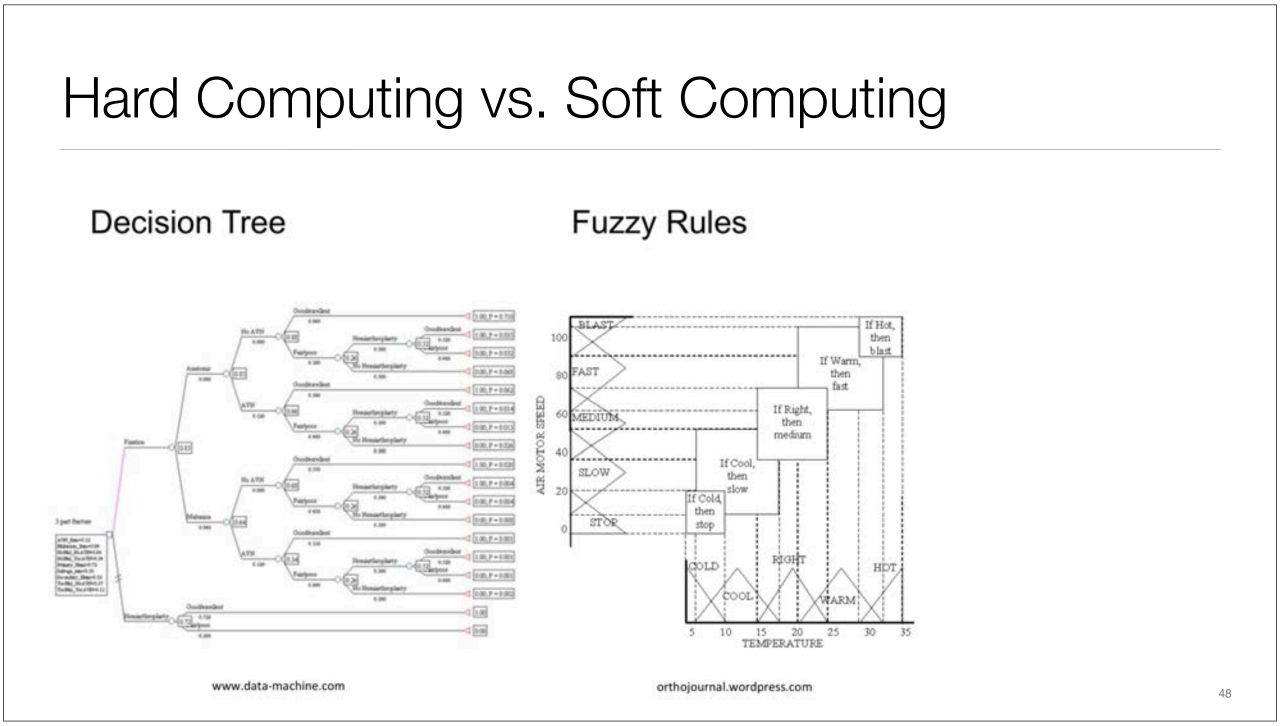
Large scale Forensic Investigations

- Situation-aware methods
- Quantified, measurable indicators
- Adaptive, self-organising models
- Distributed, cooperative, autonomous

Computational Intelligence

ANN: Neural Networks
FL: Fuzzy Logic
EC: Evolutionary Computation

HMM: Hidden-Markov Models
SOM: Self-organising Maps
SVM: Support-Vector Machines
GA: Genetic Algorithms
PSO: Particle Swarm Optimisation



Hard Computing vs. Soft Computing

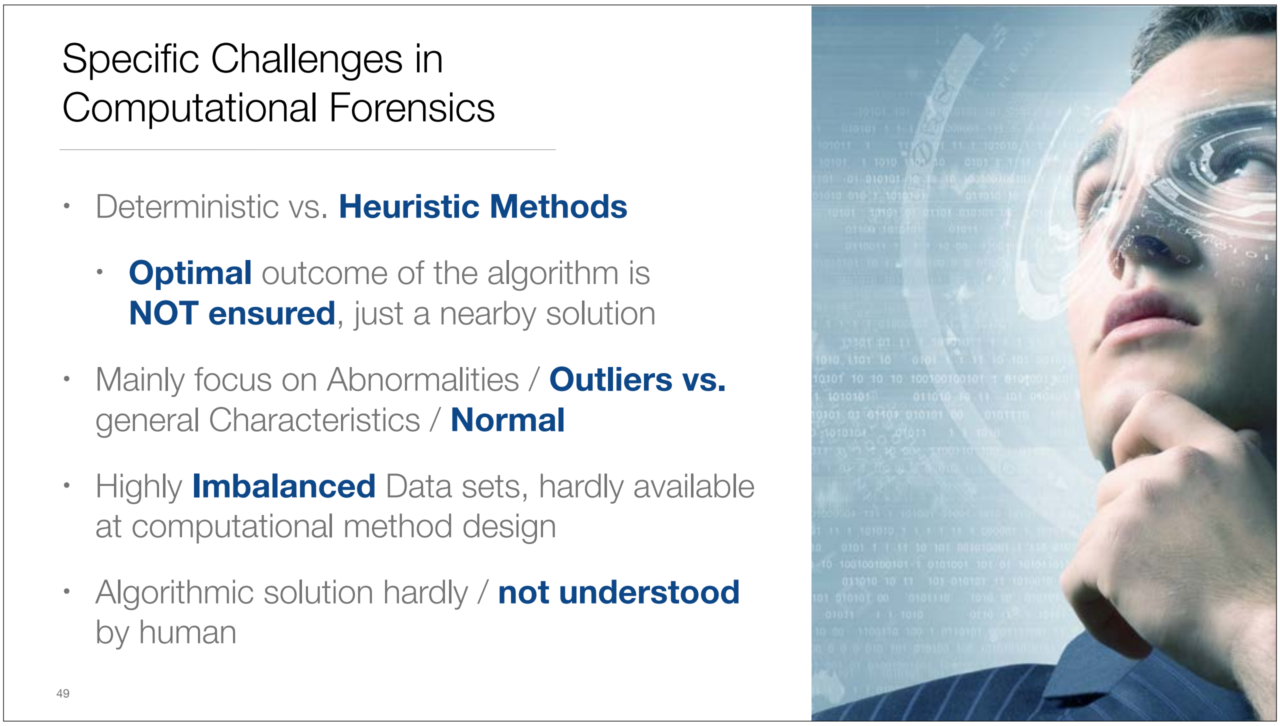
Decision Tree

Fuzzy Rules

www.data-machine.com

orthojournal.wordpress.com

This publication was realised with the EEA Financial Mechanism 2014-2021 financial support. Its content (text, photos, videos) does not reflect the official opinion of the Programme Operator, the National Contact Point and the Financial Mechanism Office. Responsibility for the information and views expressed therein lies entirely with the author(s).



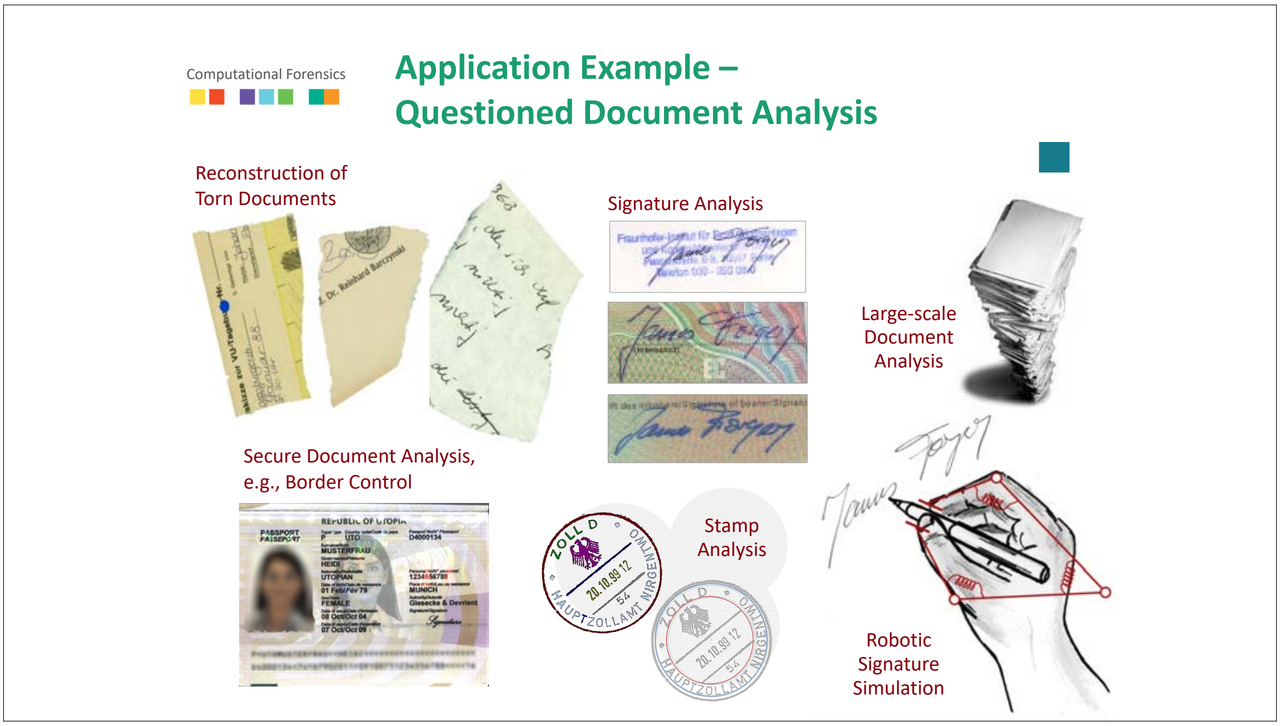
Specific Challenges in Computational Forensics

- Deterministic vs. **Heuristic Methods**
  - **Optimal** outcome of the algorithm is **NOT ensured**, just a nearby solution
- Mainly focus on Abnormalities / **Outliers vs.** general Characteristics / **Normal**
- Highly **Imbalanced** Data sets, hardly available at computational method design
- Algorithmic solution hardly / **not understood** by human

49

Economic Crime Investigation

Application Example - H2020 ESSENTIAL



Computational Forensics

Application Example – Questioned Document Analysis

Reconstruction of Torn Documents

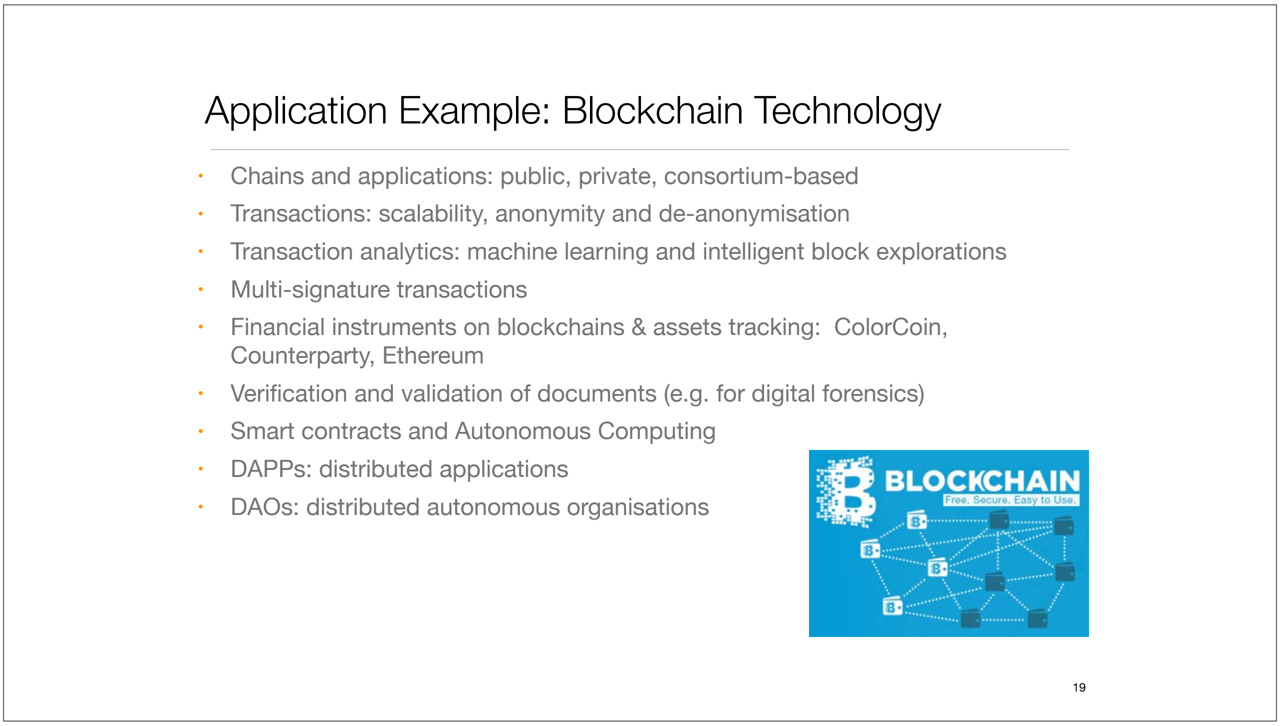
Signature Analysis

Large-scale Document Analysis

Secure Document Analysis, e.g., Border Control

Stamp Analysis

Robotic Signature Simulation



Application Example: Blockchain Technology

- Chains and applications: public, private, consortium-based
- Transactions: scalability, anonymity and de-anonymisation
- Transaction analytics: machine learning and intelligent block explorations
- Multi-signature transactions
- Financial instruments on blockchains & assets tracking: ColorCoin, Counterparty, Ethereum
- Verification and validation of documents (e.g. for digital forensics)
- Smart contracts and Autonomous Computing
- DAPPs: distributed applications
- DAOs: distributed autonomous organisations

19

This publication was realised with the EEA Financial Mechanism 2014-2021 financial support. Its content (text, photos, videos) does not reflect the official opinion of the Programme Operator, the National Contact Point and the Financial Mechanism Office. Responsibility for the information and views expressed therein lies entirely with the author(s).

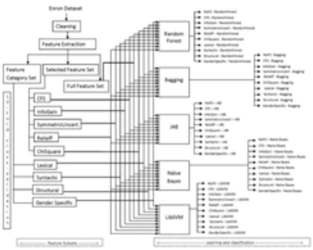
Application Example: Economic Crime Investigation

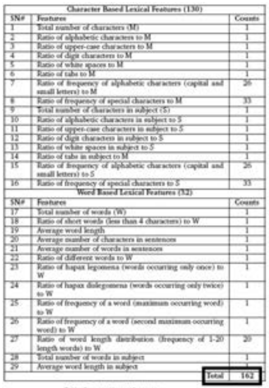
https://sites.google.com/site/drstefanaxelsson/publications

- Detection of anomalous financial and other transactions, Large quantities of data – Need automation/tools
- Self learning systems that automatically classify "unusual" behaviour or transactions:
  - These systems are opaque, the operator only sees the result, gains no insight into why the system sounded the alarm
  - Our approach, use information visualisation to make the detection system understandable by the operator – We're trying to optimise the human+machine system as a whole
- Research based on simulation of different financial systems to preserve sensitive info, allow experimentation; what-ifs, different types of fraud etc.

Application Example: Email Analysis and Author Identification from Text-based Communications

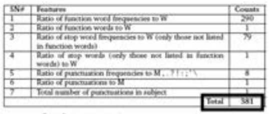
- Determining authorship of an anonymous text
- Enron dataset: real emails of Enron employees, contains **255,636 email 87,474 authors**.





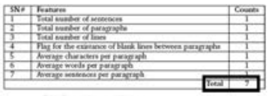
Character Based Lexical Features (130)		
SN#	Features	Counts
1	Total number of characters (M)	1
2	Ratio of alphabetic characters to M	1
3	Ratio of upper-case characters to M	1
4	Ratio of digit characters to M	1
5	Ratio of white spaces to M	1
6	Ratio of tabs to M	1
7	Ratio of frequency of alphabetic characters (capital and small letters) to M	26
8	Ratio of frequency of special characters to M	33
9	Total number of characters in subject (S)	1
10	Ratio of alphabetic characters in subject to S	1
11	Ratio of upper-case characters in subject to S	1
12	Ratio of digit characters in subject to S	1
13	Ratio of white spaces in subject to S	1
14	Ratio of tabs in subject to M	1
15	Ratio of frequency of alphabetic characters (capital and small letters) to S	26
16	Ratio of frequency of special characters to S	33
Word Based Lexical Features (32)		
SN#	Features	Counts
17	Total number of words (W)	1
18	Ratio of short words (less than 4 characters) to W	1
19	Average word length	1
20	Average number of characters in sentences	1
21	Average number of words in sentences	1
22	Ratio of different words to W	1
23	Ratio of hapax legomena (words occurring only once) to W	1
24	Ratio of hapax dislegomena (words occurring only twice) to W	1
25	Ratio of frequency of a word (maximum occurring word) to W	1
26	Ratio of frequency of a word (second maximum occurring word) to W	1
27	Ratio of word length distribution (frequency of 1-20 length words) to W	20
28	Total number of words in subject	1
29	Average word length in subject	1
	Total	162

Table: Extracted lexical features



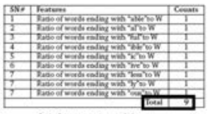
SN#	Features	Counts
1	Ratio of function word frequencies to W	290
2	Ratio of function words to W	1
3	Ratio of stop word frequencies to W (only those not listed in function words)	79
4	Ratio of stop words (only those not listed in function words) to W	1
5	Ratio of punctuation frequencies to M , . ? ! : ; ' "	8
6	Ratio of punctuations to M	1
7	Total number of punctuations in subject	1
	Total	381

Table: Extracted syntactic features



SN#	Features	Counts
1	Total number of sentences	1
2	Total number of paragraphs	1
3	Total number of lines	1
4	Flag for the existance of blank lines between paragraphs	1
5	Average characters per paragraph	1
6	Average words per paragraph	1
7	Average sentences per paragraph	1
	Total	7

Table: Extracted structural features



SN#	Features	Counts
1	Ratio of words ending with "able"to W	1
2	Ratio of words ending with "al"to W	1
3	Ratio of words ending with "ful"to W	1
4	Ratio of words ending with "ible"to W	1
5	Ratio of words ending with "ic"to W	1
6	Ratio of words ending with "ive"to W	1
7	Ratio of words ending with "less"to W	1
7	Ratio of words ending with "ly"to W	1
7	Ratio of words ending with "ous"to W	1
	Total	9

Table: Extracted gender specific features

Reference: Chitrakar, Norbø, Franke (2011-)

This publication was realised with the EEA Financial Mechanism 2014-2021 financial support. Its content (text, photos, videos) does not reflect the official opinion of the Programme Operator, the National Contact Point and the Financial Mechanism Office. Responsibility for the information and views expressed therein lies entirely with the author(s).

Analysis by Synthesis

Our current domain: Financial Fraud Research / Tax Evasion / Money Laundering

- Leaning from real-world data with restricted access
- Privacy of customers is not affected

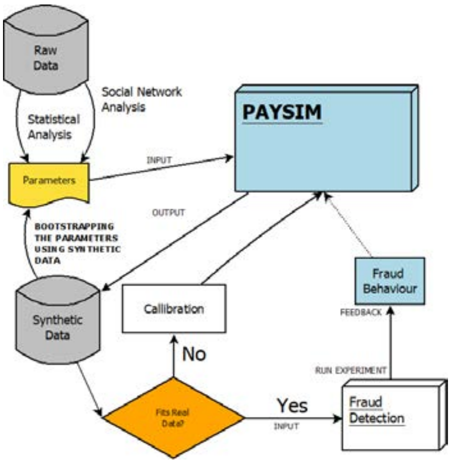
- Results can be disclosed to, and compared by, other researchers
- Different scenarios can be modelled using well controlled parameters

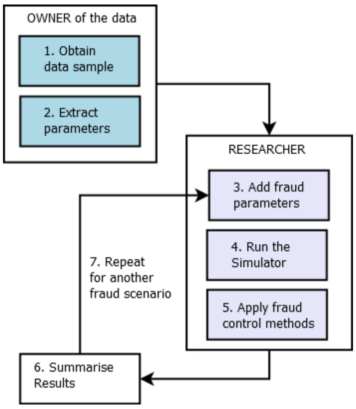
- Avoid some of the Machine Learning challenges , i.e. Class-Imbalance, non labelled data

- Use it for Training non experts in a field to become familiar with diverse scenarios before they ever seen it

Financial Fraud Simulator

E.A. Lopez: https://tinyurl.com/yaoll8fk
Our data: https://www.kaggle.com/ntnu-testimon/banksim1





This publication was realised with the EEA Financial Mechanism 2014-2021 financial support. Its content (text, photos, videos) does not reflect the official opinion of the Programme Operator, the National Contact Point and the Financial Mechanism Office. Responsibility for the information and views expressed therein lies entirely with the author(s).

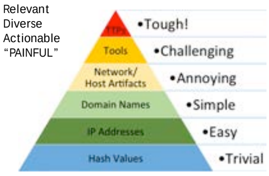
Threat Intelligence, Information Fusion & Sharing

Application Example - BIA ACT

Application Example: Cyber Threat Intelligence

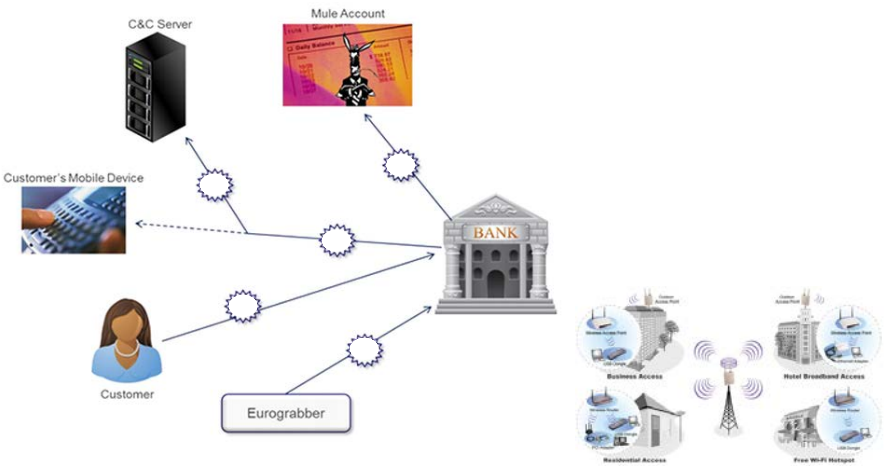
- Gartner's definition:
  - "***Evidence-based knowledge***, including context, mechanisms, indicators, implications and actionable advice, about an ***existing or emerging menace or hazard*** to assets that can be ***used to inform decisions*** regarding the subject's response to that menace or hazard"
- Proactive Cyber Security
  - Research on: Tactics, Techniques, and Procedures
  - Understand security trends and risks
- Sources of ThreatIntel
  - Private Commercial Providers
  - Public (e.g. government security institutions)
  - Malware analysis reports and feeds
  - Incident reports
  - Vendors reports
  - Open sources (e.g. social media, news, blogs)
  - "Hacker Forums"
  - Use to share/trade/exchange hacking services, tools, etc.

- The properties of "good" intelligence:
  - Accurate
  - Relevant
  - Diverse
  - Actionable
  - "PAINFUL"



This publication was realised with the EEA Financial Mechanism 2014-2021 financial support. Its content (text, photos, videos) does not reflect the official opinion of the Programme Operator, the National Contact Point and the Financial Mechanism Office. Responsibility for the information and views expressed therein lies entirely with the author(s).

ACT - Project Example



POLITIET KRIPOS

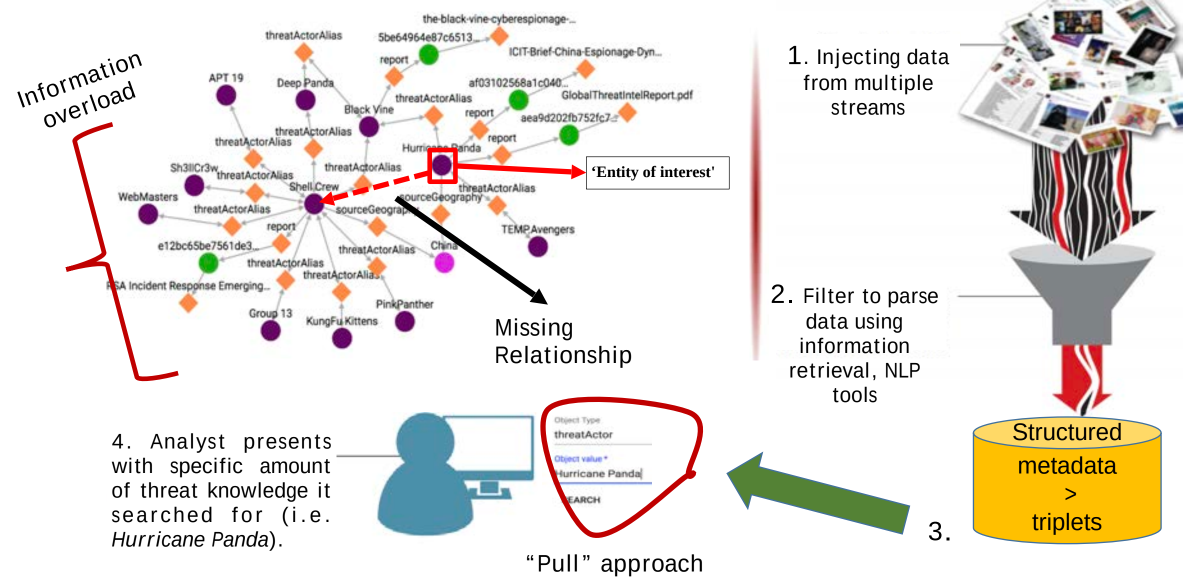
FinansCERT

mnemonic

NTNU

CCIS

Application Example: Threat Intelligent Platforms



Information overload

Missing Relationship

1. Injecting data from multiple streams

2. Filter to parse data using information retrieval, NLP tools

Structured metadata > triplets

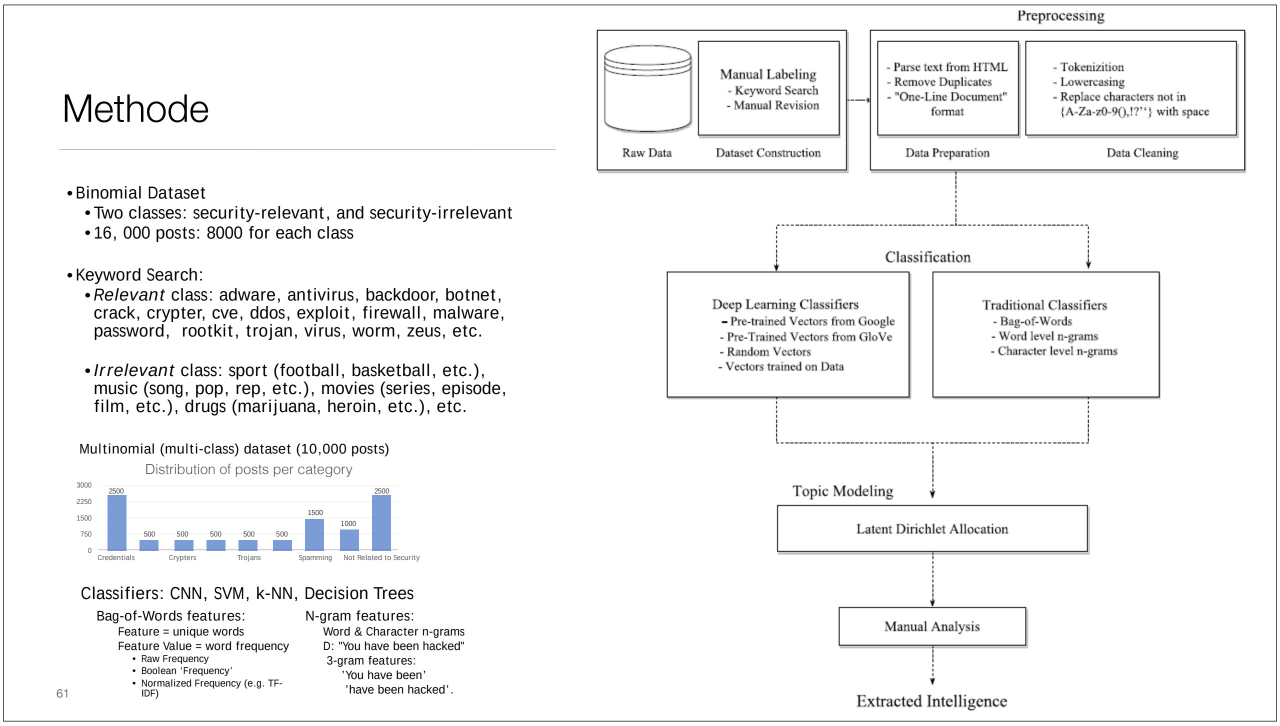
3.

4. Analyst presents with specific amount of threat knowledge it searched for (i.e. *Hurricane Panda*).

Object Type
threatActor
Object value *
Hurricane Panda
SEARCH

"Pull" approach

This publication was realised with the EEA Financial Mechanism 2014-2021 financial support. Its content (text, photos, videos) does not reflect the official opinion of the Programme Operator, the National Contact Point and the Financial Mechanism Office. Responsibility for the information and views expressed therein lies entirely with the author(s).



Methode

- Binomial Dataset
  - Two classes: security-relevant, and security-irrelevant
  - 16, 000 posts: 8000 for each class
- Keyword Search:
  - *Relevant* class: adware, antivirus, backdoor, botnet, crack, crypter, cve, ddos, exploit, firewall, malware, password, rootkit, trojan, virus, worm, zeus, etc.
  - *Irrelevant* class: sport (football, basketball, etc.), music (song, pop, rep, etc.), movies (series, episode, film, etc.), drugs (marijuana, heroin, etc.), etc.

Multinomial (multi-class) dataset (10,000 posts)

Classifiers: CNN, SVM, k-NN, Decision Trees

Bag-of-Words features:
Feature = unique words
Feature Value = word frequency
- Raw Frequency
- Boolean ‘Frequency’
- Normalized Frequency (e.g. TF-IDF)

N-gram features:
Word & Character n-grams
D: "You have been hacked"
3-gram features:
’You have been’
’have been hacked’.

61

Preprocessing

Raw Data

Manual Labeling
- Keyword Search
- Manual Revision

Dataset Construction

- Parse text from HTML
- Remove Duplicates
- "One-Line Document" format

Data Preparation

- Tokenizition
- Lowercasing
- Replace characters not in {A-Za-z0-9(),!?’‘} with space

Data Cleaning

Classification

Deep Learning Classifiers
- Pre-trained Vectors from Google
- Pre-Trained Vectors from GloVe
- Random Vectors
- Vectors trained on Data

Traditional Classifiers
- Bag-of-Words
- Word level n-grams
- Character level n-grams

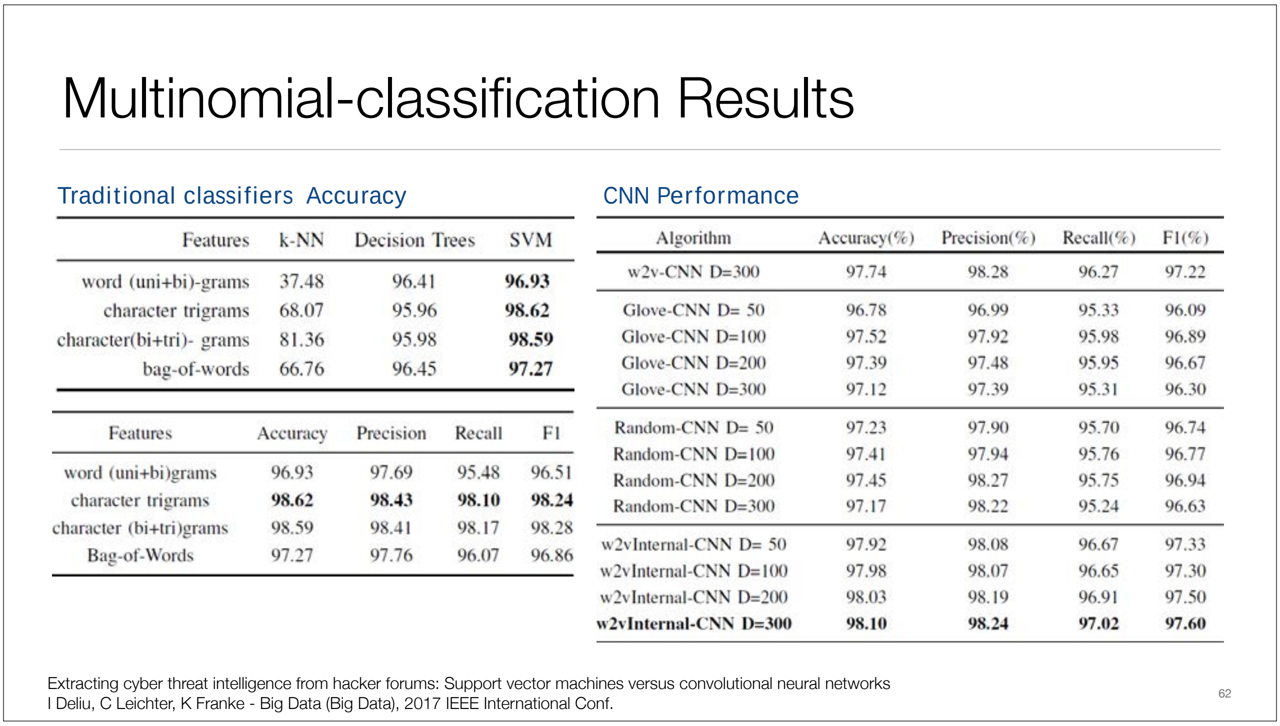
Topic Modeling

Latent Dirichlet Allocation

Manual Analysis

Extracted Intelligence

61



Multinomial-classification Results

Traditional classifiers Accuracy

Features	k-NN	Decision Trees	SVM
word (uni+bi)-grams	37.48	96.41	**96.93**
character trigrams	68.07	95.96	**98.62**
character(bi+tri)- grams	81.36	95.98	**98.59**
bag-of-words	66.76	96.45	**97.27**

Features	Accuracy	Precision	Recall	F1
word (uni+bi)grams	96.93	97.69	95.48	96.51
character trigrams	**98.62**	**98.43**	**98.10**	**98.24**
character (bi+tri)grams	98.59	98.41	98.17	98.28
Bag-of-Words	97.27	97.76	96.07	96.86

CNN Performance

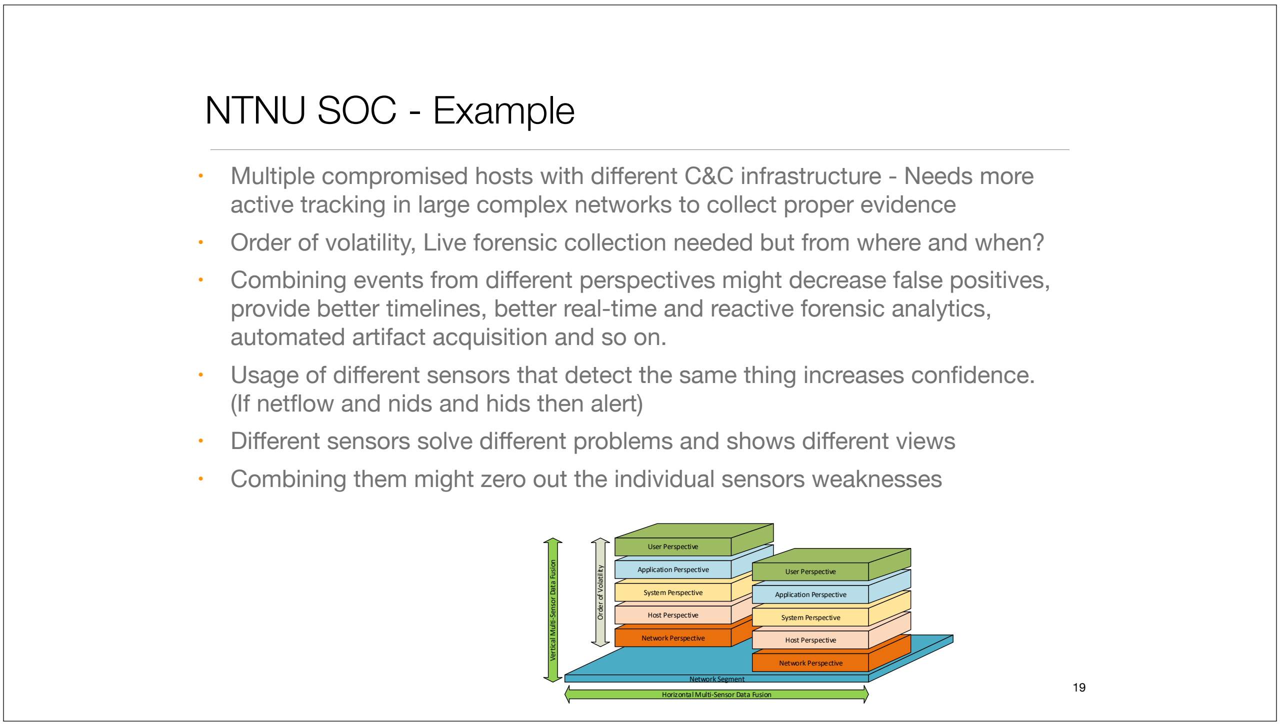
Algorithm	Accuracy(%)	Precision(%)	Recall(%)	F1(%)
w2v-CNN D=300	97.74	98.28	96.27	97.22
Glove-CNN D= 50	96.78	96.99	95.33	96.09
Glove-CNN D=100	97.52	97.92	95.98	96.89
Glove-CNN D=200	97.39	97.48	95.95	96.67
Glove-CNN D=300	97.12	97.39	95.31	96.30
Random-CNN D= 50	97.23	97.90	95.70	96.74
Random-CNN D=100	97.41	97.94	95.76	96.77
Random-CNN D=200	97.45	98.27	95.75	96.94
Random-CNN D=300	97.17	98.22	95.24	96.63
w2vInternal-CNN D= 50	97.92	98.08	96.67	97.33
w2vInternal-CNN D=100	97.98	98.07	96.65	97.30
w2vInternal-CNN D=200	98.03	98.19	96.91	97.50
w2vInternal-CNN D=300	**98.10**	**98.24**	**97.02**	**97.60**

Extracting cyber threat intelligence from hacker forums: Support vector machines versus convolutional neural networks
I Deliu, C Leichter, K Franke - Big Data (Big Data), 2017 IEEE International Conf.

62

62

This publication was realised with the EEA Financial Mechanism 2014-2021 financial support. Its content (text, photos, videos) does not reflect the official opinion of the Programme Operator, the National Contact Point and the Financial Mechanism Office. Responsibility for the information and views expressed therein lies entirely with the author(s).



NTNU SOC - Example

- Multiple compromised hosts with different C&C infrastructure - Needs more active tracking in large complex networks to collect proper evidence
- Order of volatility, Live forensic collection needed but from where and when?
- Combining events from different perspectives might decrease false positives, provide better timelines, better real-time and reactive forensic analytics, automated artifact acquisition and so on.
- Usage of different sensors that detect the same thing increases confidence. (If netflow and nids and hids then alert)
- Different sensors solve different problems and shows different views
- Combining them might zero out the individual sensors weaknesses

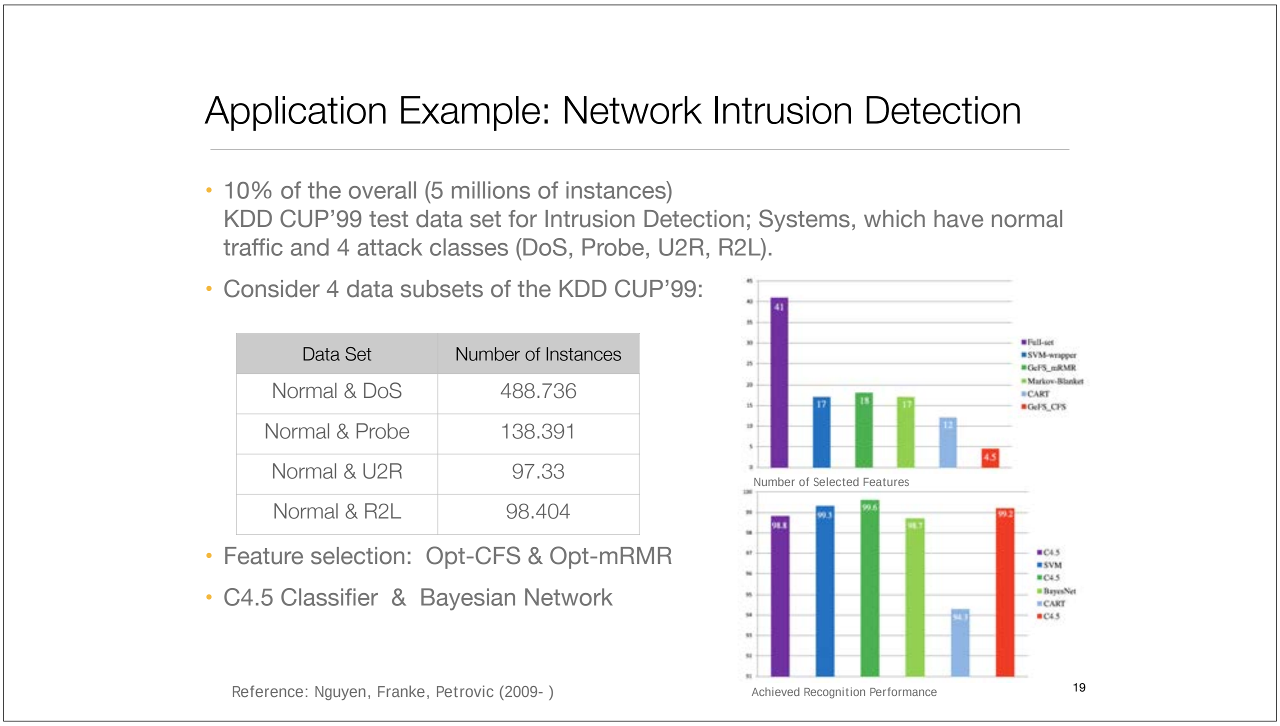
19

63

Malicious Code Detection

Application Example - NFR Ars Forensica

64



Application Example: Network Intrusion Detection

- 10% of the overall (5 millions of instances) KDD CUP'99 test data set for Intrusion Detection; Systems, which have normal traffic and 4 attack classes (DoS, Probe, U2R, R2L).
- Consider 4 data subsets of the KDD CUP'99:

Data Set	Number of Instances
Normal & DoS	488.736
Normal & Probe	138.391
Normal & U2R	97.33
Normal & R2L	98.404

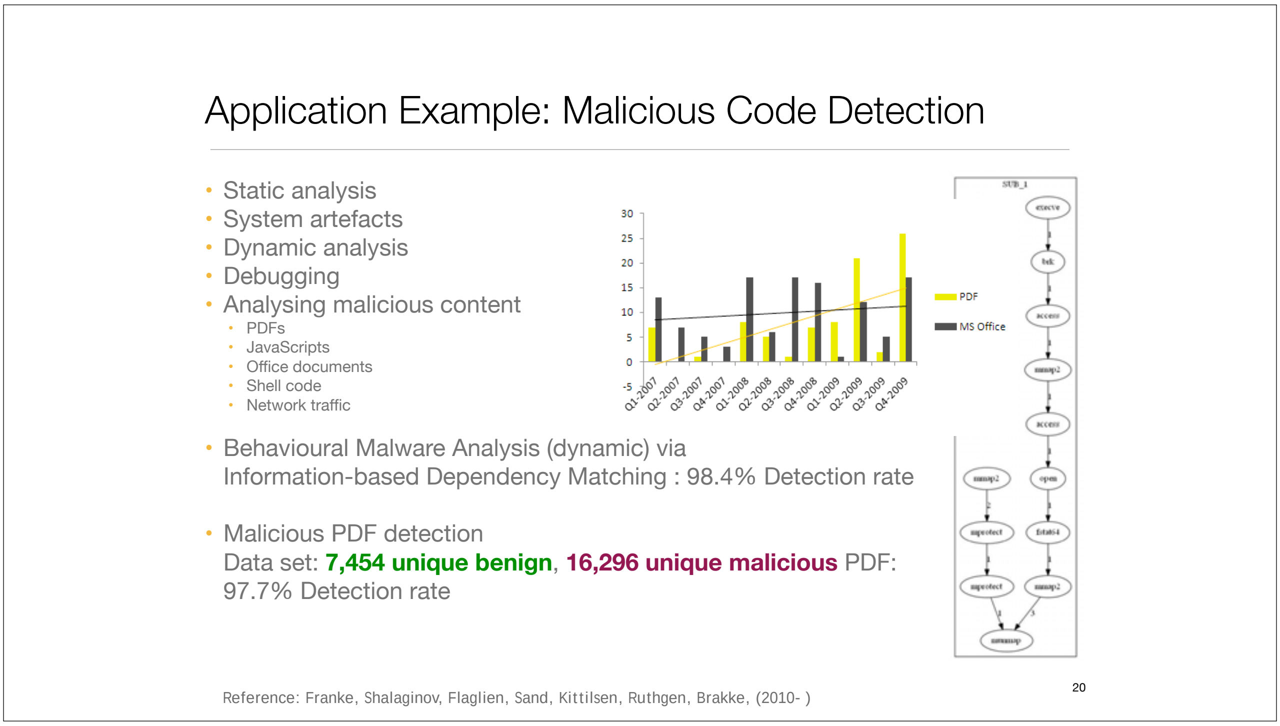
- Feature selection: Opt-CFS & Opt-mRMR
- C4.5 Classifier & Bayesian Network

Number of Selected Features

Achieved Recognition Performance

Reference: Nguyen, Franke, Petrovic (2009-)

19



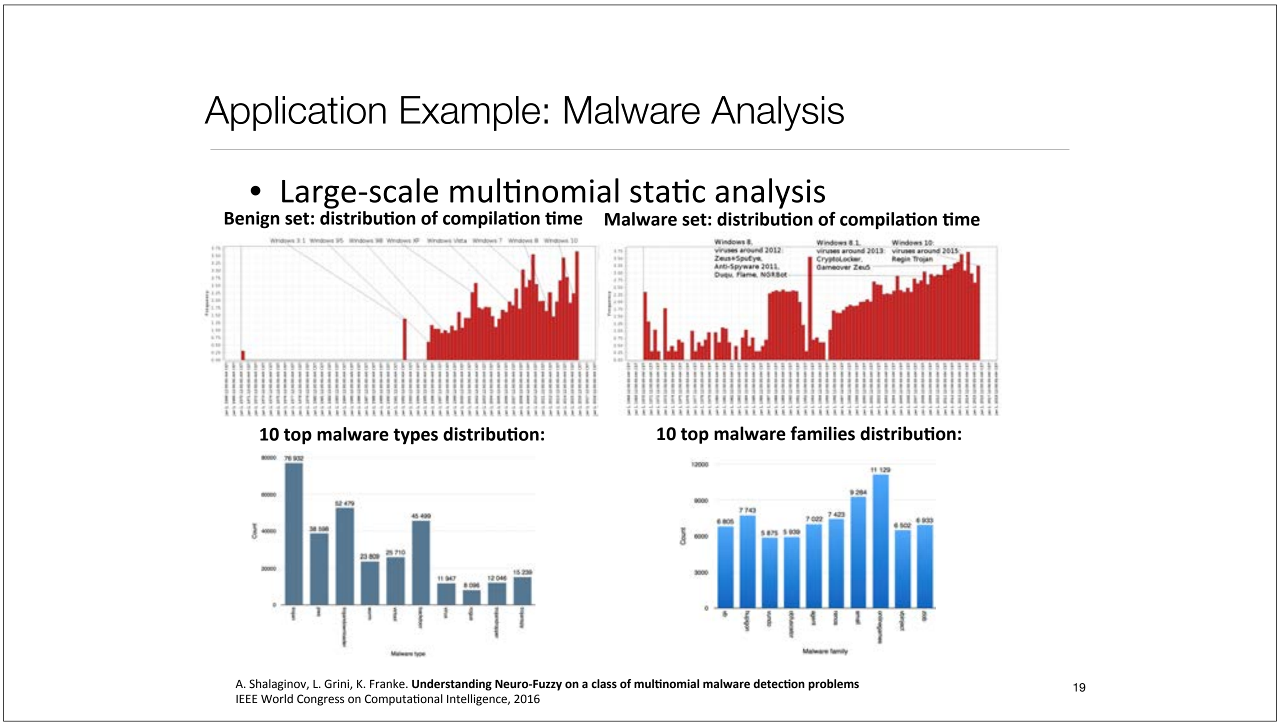
Application Example: Malicious Code Detection

- Static analysis
- System artefacts
- Dynamic analysis
- Debugging
- Analysing malicious content
  - PDFs
  - JavaScripts
  - Office documents
  - Shell code
  - Network traffic
- Behavioural Malware Analysis (dynamic) via Information-based Dependency Matching : 98.4% Detection rate
- Malicious PDF detection
  Data set: **7,454 unique benign**, **16,296 unique malicious** PDF: 97.7% Detection rate

Reference: Franke, Shalaginov, Flaglien, Sand, Kittilsen, Ruthgen, Brakke, (2010-)

20

This publication was realised with the EEA Financial Mechanism 2014-2021 financial support. Its content (text, photos, videos) does not reflect the official opinion of the Programme Operator, the National Contact Point and the Financial Mechanism Office. Responsibility for the information and views expressed therein lies entirely with the author(s).



Application Example: Malware Analysis

- Large-scale multinomial static analysis

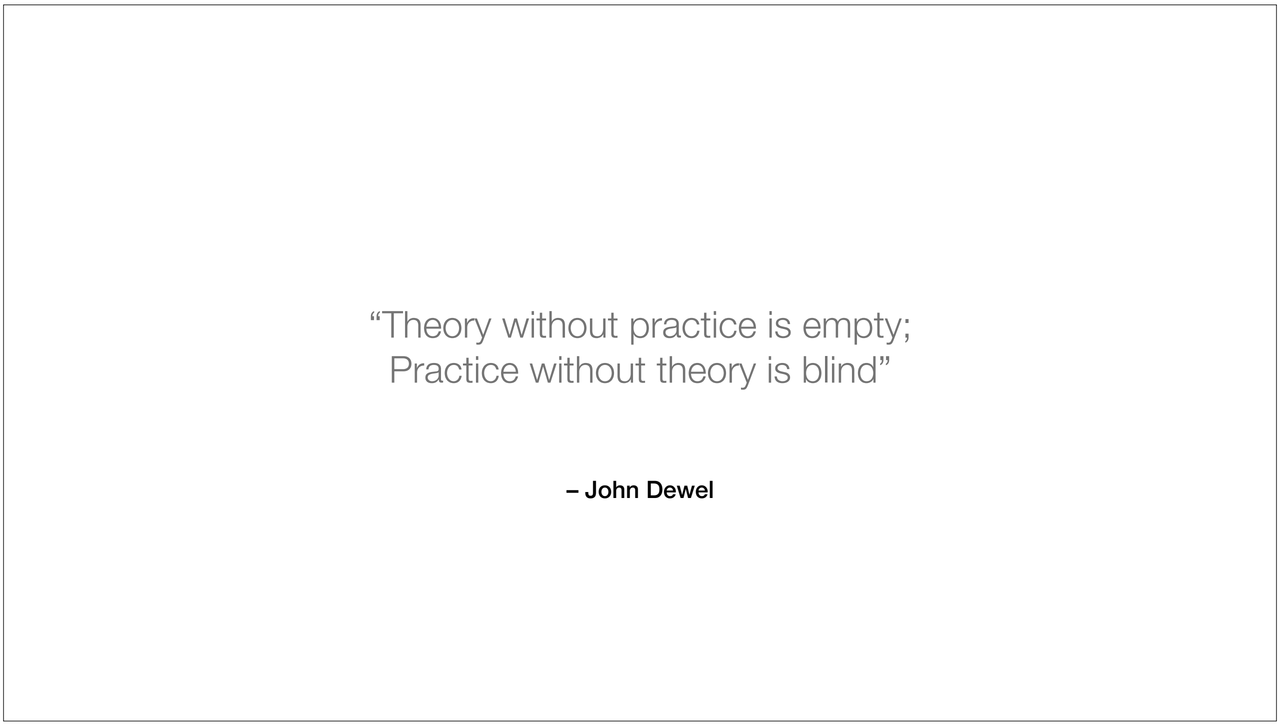
Benign set: distribution of compilation time **Malware set: distribution of compilation time**

10 top malware types distribution: **10 top malware families distribution:**

A. Shalaginov, L. Grini, K. Franke. **Understanding Neuro-Fuzzy on a class of multinomial malware detection problems**
IEEE World Congress on Computational Intelligence, 2016

19

67

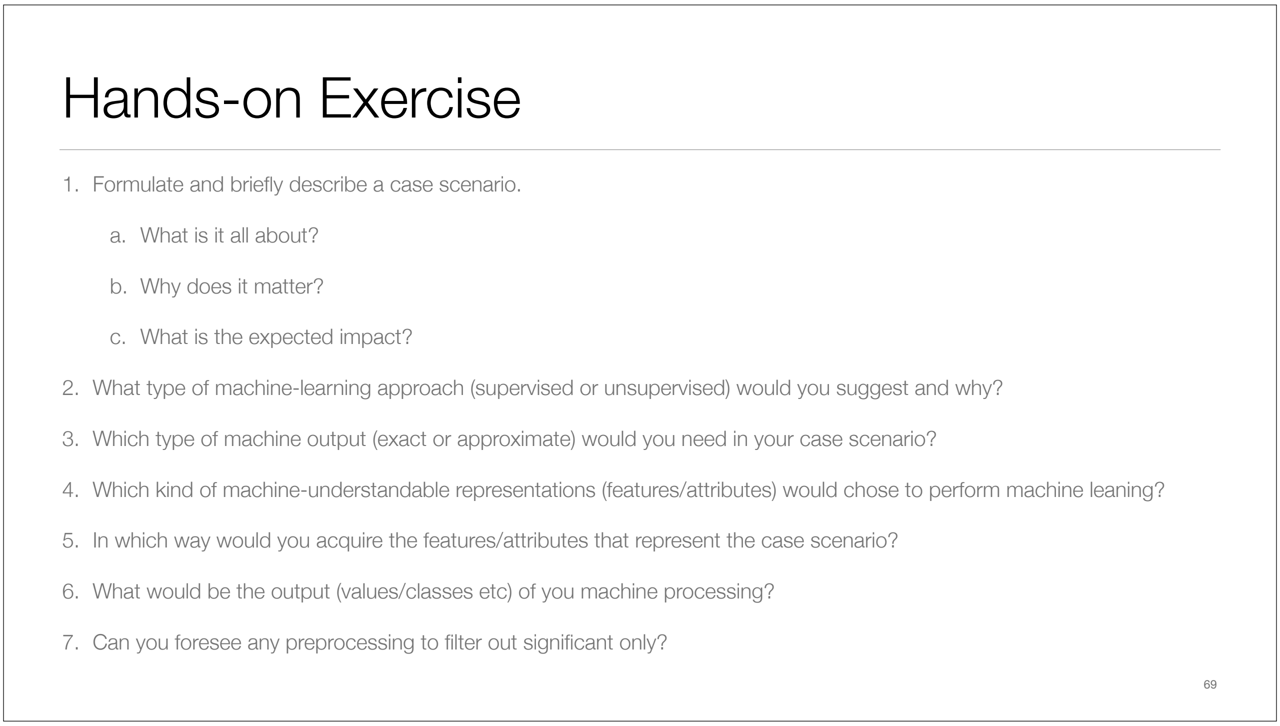


“Theory without practice is empty;
Practice without theory is blind”

– John Dewel

68

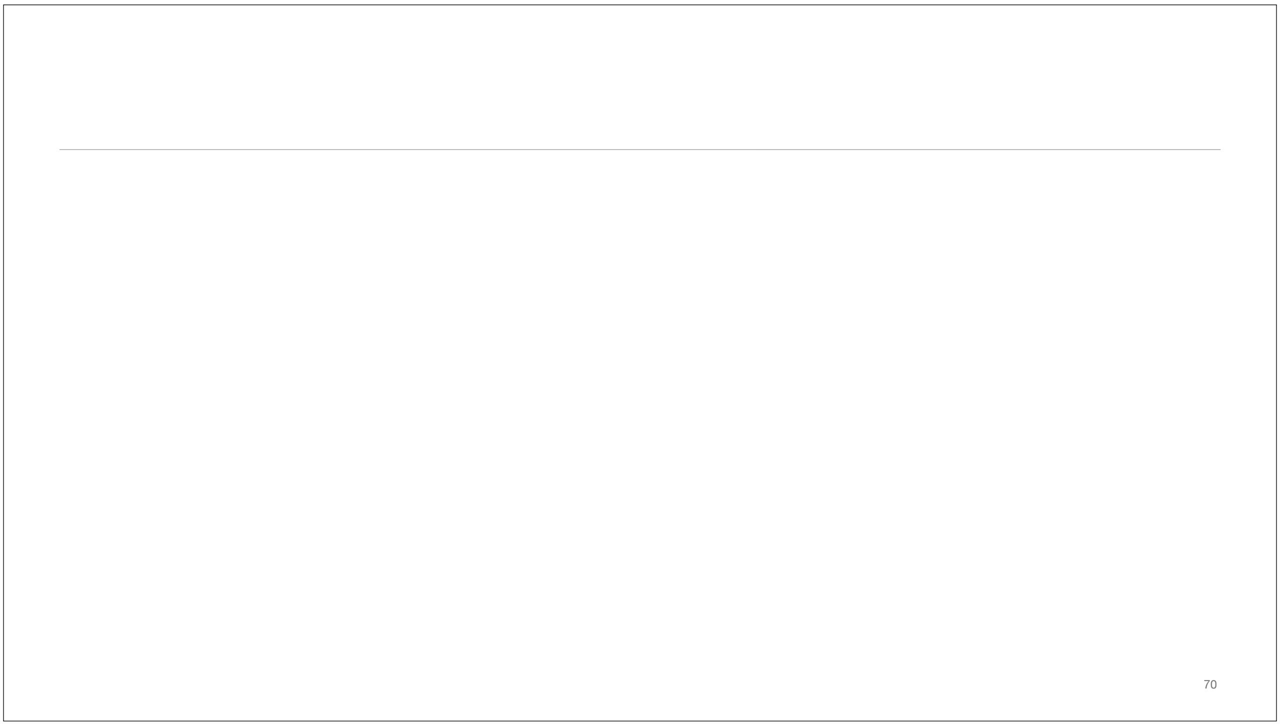
This publication was realised with the EEA Financial Mechanism 2014-2021 financial support. Its content (text, photos, videos) does not reflect the official opinion of the Programme Operator, the National Contact Point and the Financial Mechanism Office. Responsibility for the information and views expressed therein lies entirely with the author(s).



Hands-on Exercise

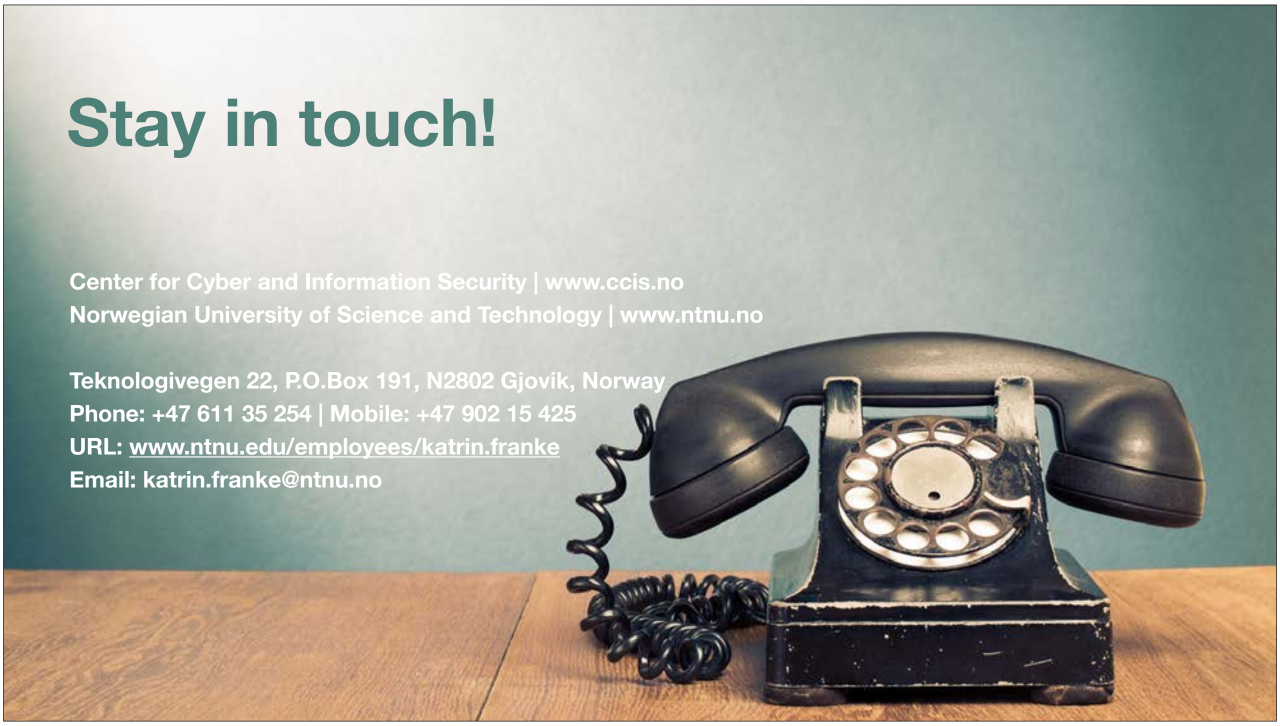
1. Formulate and briefly describe a case scenario.
    a. What is it all about?
    b. Why does it matter?
    c. What is the expected impact?
2. What type of machine-learning approach (supervised or unsupervised) would you suggest and why?
3. Which type of machine output (exact or approximate) would you need in your case scenario?
4. Which kind of machine-understandable representations (features/attributes) would chose to perform machine leaning?
5. In which way would you acquire the features/attributes that represent the case scenario?
6. What would be the output (values/classes etc) of you machine processing?
7. Can you foresee any preprocessing to filter out significant only?

69



70

This publication was realised with the EEA Financial Mechanism 2014-2021 financial support. Its content (text, photos, videos) does not reflect the official opinion of the Programme Operator, the National Contact Point and the Financial Mechanism Office. Responsibility for the information and views expressed therein lies entirely with the author(s).

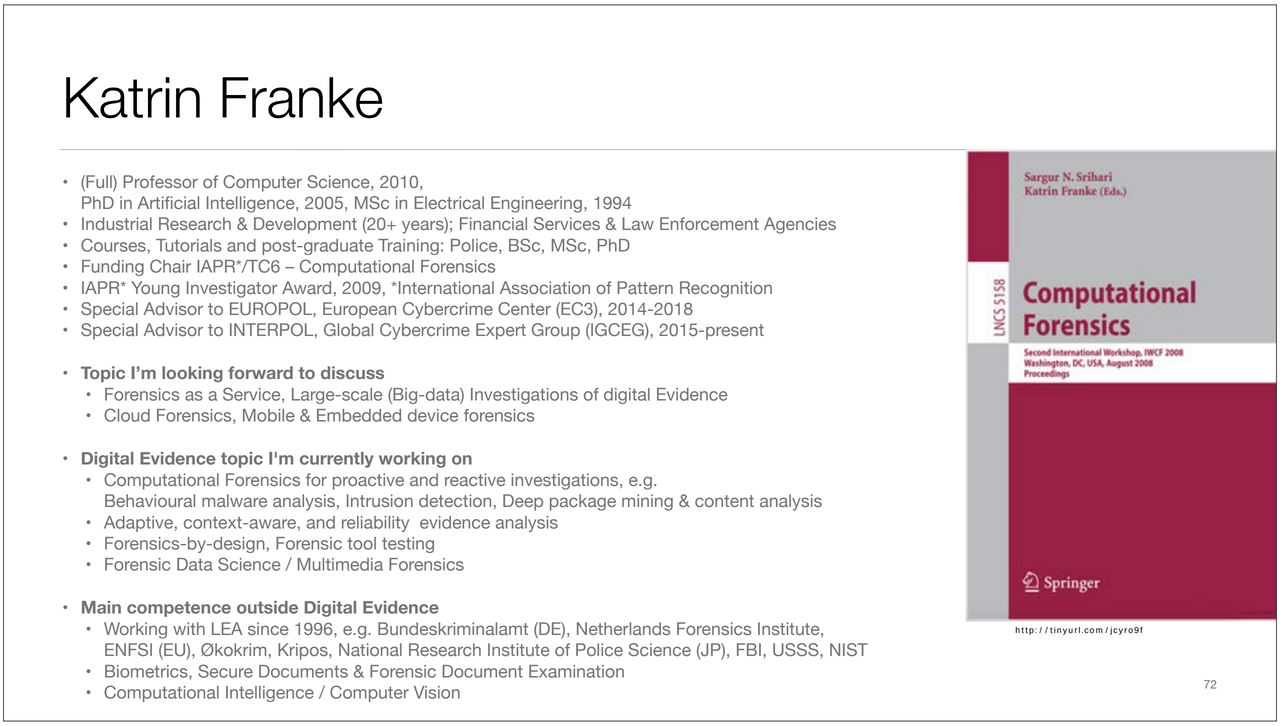


Stay in touch!

Center for Cyber and Information Security | www.ccis.no
Norwegian University of Science and Technology | www.ntnu.no

Teknologivegen 22, P.O.Box 191, N2802 Gjovik, Norway
Phone: +47 611 35 254 | Mobile: +47 902 15 425
URL: www.ntnu.edu/employees/katrin.franke
Email: katrin.franke@ntnu.no

71



Katrin Franke

- (Full) Professor of Computer Science, 2010,
  PhD in Artificial Intelligence, 2005, MSc in Electrical Engineering, 1994
- Industrial Research & Development (20+ years); Financial Services & Law Enforcement Agencies
- Courses, Tutorials and post-graduate Training: Police, BSc, MSc, PhD
- Funding Chair IAPR*/TC6 – Computational Forensics
- IAPR* Young Investigator Award, 2009, *International Association of Pattern Recognition
- Special Advisor to EUROPOL, European Cybercrime Center (EC3), 2014-2018
- Special Advisor to INTERPOL, Global Cybercrime Expert Group (IGCEG), 2015-present

- **Topic I'm looking forward to discuss**
  - Forensics as a Service, Large-scale (Big-data) Investigations of digital Evidence
  - Cloud Forensics, Mobile & Embedded device forensics

- **Digital Evidence topic I'm currently working on**
  - Computational Forensics for proactive and reactive investigations, e.g.
    Behavioural malware analysis, Intrusion detection, Deep package mining & content analysis
  - Adaptive, context-aware, and reliability evidence analysis
  - Forensics-by-design, Forensic tool testing
  - Forensic Data Science / Multimedia Forensics

- **Main competence outside Digital Evidence**
  - Working with LEA since 1996, e.g. Bundeskriminalamt (DE), Netherlands Forensics Institute,
    ENFSI (EU), Økokrim, Kripos, National Research Institute of Police Science (JP), FBI, USSS, NIST
  - Biometrics, Secure Documents & Forensic Document Examination
  - Computational Intelligence / Computer Vision

http://tinyurl.com/jcyro9f

72

72