NTNU CCIS

Computational Thinking

Combining Police Intelligence & AI

Katrin FRANKE, PhD Professor of Computer Science

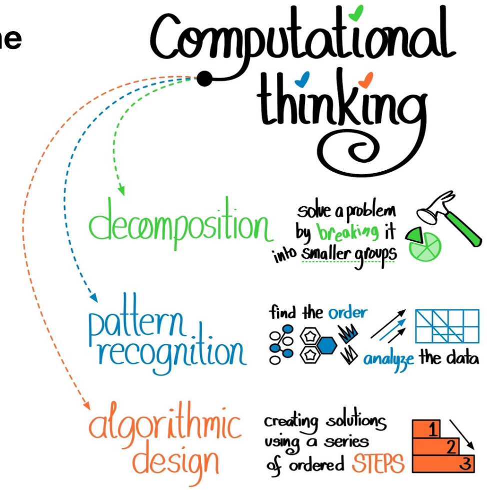
Center for Cyber and Information Security | www.ccis.no
Norwegian University of Science and Technology | www.ntnu.no

1

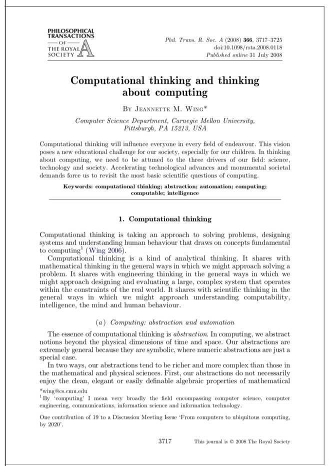
2

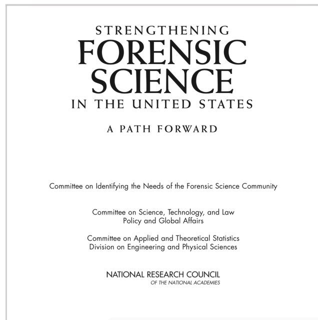
A digital age skill for everyone

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



3





4

Computational Forensics

- Study and development of computational methods to
  - **Assist in basic and applied research**, e.g. to **establish or prove the scientific basis** of a particular investigative procedure,
  - **Support the forensic examiner** in their daily casework.
- Modern crime investigation shall profit from the hybrid-intelligence of humans and machines.

5

Three Professorship in DF (since 2014)

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

6

6

NTNU Digital Forensics Group @IIK

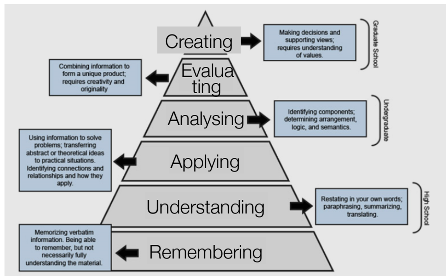
- 1+3 (Assoc.) Professors, 4+1 Postdoc, 15+3 PhD Students, 5 Adjunct Researchers, 1 Project Admin, ca. 20 Master Students per year, 3 Professors **financed by the Police directorate**
- **1 Focus - Technological aspects of digital & computational forensics**
  Teaching on Bachelor, Master, and PhD Level; 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, 2015-2019
  HANSKEN - **Norwegian Police**, Big Data Forensics: Infrastructure, 2016-2018
  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

7

Education & Training

- Tasks require different **Knowledge, Skills, and General Competences**
- Education and Training shall address **different demands**, i.e. First Responder vs. Special Investigator
- **Continuous Learning** and Adoption of new knowledge and skills is required
- **Research-based Education** to follow / be at the forefront of technology development
- BSc, MSc, & PhD Level Education



Bloom's taxonomy
Classification system of educational objectives *(Version 2001)*

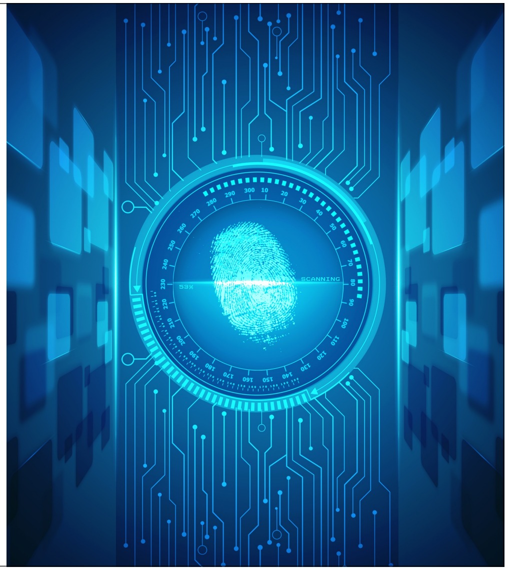
8

8

Research Agenda

- Computational Forensics
  - Reliable Algorithms
  - Forensic as a Service using secure Computing infrastructure
- Cloud Forensics & Cybercrime Investigation
  - *Sergii Bian - DFRWS '18*
  - *Kyle Porter - DFRWS '18*
- Economic Crime Investigation
- Mobile & Embedded Device Forensics (IoT, IoE)
  - *Gunnar Alendahl - DFRWS-EU '18*
  - *Jens-Petter Sandvik - DFRWS-EU '18*

9



9

Perspectives on Digital Investigation

- **Legal** / Regulations / Policies / Rule of Law
- ★ **Technological** / Security / Archival
- **Organisational** / Information Management / Procedures / Governance
- **Knowledge** / Capacity Building / Training Public Awareness (pedagogical methods)

10

10

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
- Computational Forensics, aka applying **Computational Intelligence in Forensic Sciences**

11

11

Scenarios of Large-Scale Investigations in LEA

- **Many conventional cases** (murder, robbery, etc), e.g. Regional Police District *(Oslo)*
  - Many small data seizures can add up to
  - Several TB of data stored as evidence
  - Analysis for each case is not complex
  - Prefer analysis interface directly with front line investigators
- **Few unconventional cases**, e.g. Economic-crime Unit *(ØKOKRIM)*
  - A single case can result in large data seizures equal to many TB
  - Millions of documents, Hard drives, mobile devices
  - Analysis for each case can take years
- **Both Scenarios** => **Many TBs of Data** => **Computational Analysis**

12

12

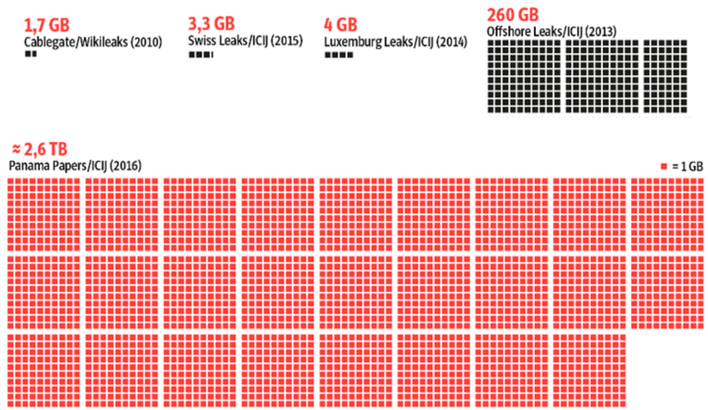
Case Scenarios: Economic-crime Unit

- **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

13

13

Panama Papers in Size Perspective



14

14

International Case Statistics

- **Normal** for cases under 100,000 documents;
- **Large** for cases with 100,000 to 1 million documents;
- **Very Large** for cases between 1 million and 100 million documents; and
- **Ridiculous**, reserved for cases with greater than 100 million documents.

LJN LAW JOURNAL NEWSLETTERS
LEGAL TECH NEWSLETTER
JANUARY 2016
Defining Big Data In the e-Discovery World
By **John Ferguson** (/authors/885.html)

Across the "Relativity universe", separate percentages are tracked for each grouping. Assessing the percentages over the past five years reveals that approximately

- **two thirds** of cases fall in the ***Normal*** group,
- approximately **a quarter** of cases in the ***Large*** group, and
- around **8%** in the ***Very Large*** group.

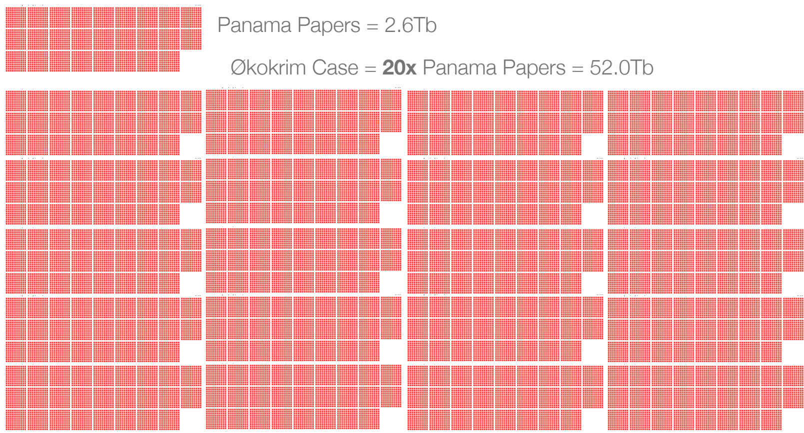
These percentages have held fairly constant over the past five years with the exception of the ***Ridiculous*** cases, which first appeared in 2013, and now, while increasing, account for **less than 1%** of the overall case size make up

Source: © kCura - Manufacturer of Relativity One of the Leading E-Discovery Tools

15

15

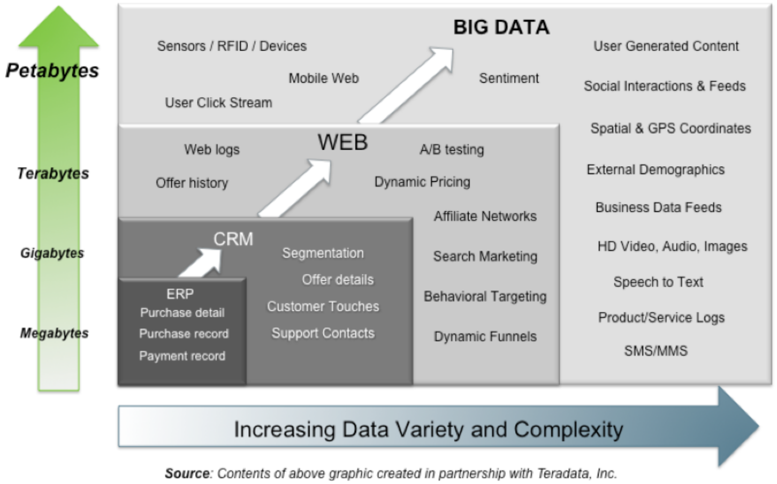
Økokrim Largest Ongoing Investigation



16

16

Big Data = Transactions + Interactions + Observations



Source: *Contents of above graphic created in partnership with Teradata, Inc.*

17

Computational Forensics

Scientific Computing in Forensics

Definitions

Forensic Science

- an applied natural science
- work to serve and provide the investigatory methods, i.e. scientific methods, in order to solve the specific crimes / accidents
- provide evidences, which are used in criminology
- based in the vast and deep studies of research, e.g. *biology*, *chemistry*, *finance*, *computing*, etc
- does not develop theories and thesis regarding any crime

Criminology

- specialised social science, which evolves from sociology
- a scientific study of nature, extent, causes, control, and prevention of the criminal behaviour of both the individual and society
- provide the criminal profile by studying the crimes and nature of the criminals
- based on the three theories: *Classical*, *Positive*, and *Chicago*
- do develop theories and thesis from their research and experience

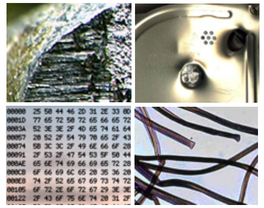
19

http://www.differencebetween.info/difference-between-forensic-science-and-criminology

Challenges & Demands in Forensic Investigations

Challenges

- **Tiny Pieces of Evidence** are hidden in a mostly **Chaotic Environment**,
- Trace Study to reveal **Specific Properties**,
- Traces found will be **Never Identical**,
- Reasoning and Deduction have to be performed on the basis of
  - **Partial Knowledge**,
  - **Approximations**,
  - **Uncertainties** and
  - **Conjectures**.



Demands

- **Objective Measurement** and **Classification**,
- **Robustness** and **Reproducibility**,
- **Secure against Falsifications**.

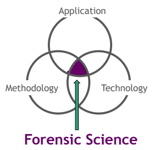
20

Computational Forensics - Definition

It is understood as the hypothesis-driven investigation of a specific forensic problem using computers, with the primary goal of discovery and advancement of forensic knowledge.

CF works towards:

1. **In-depth Understanding** of a forensic discipline,
2. **Evaluation** of a particular scientific method basis and
3. **Systematic Approach** to forensic sciences by applying techniques of computer science, applied mathematics and statistics.



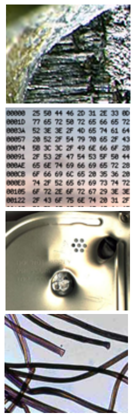
It involves **Modelling** and computer **Simulation (Synthesis)** and/or computer-based **Analysis** and **Recognition**

Computational Methods

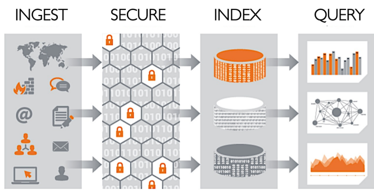
- **Signal / Image Processing** : one-dimensional signals and two-dimensional images are transformed for better human or machine processing,
- **Computer Vision** : images are automatically recognised to identify objects,
- **Computer Graphics / Data Visualisation** : two-dimensional images or three-dimensional 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.

Computational vs. Computer (Digital) Forensics

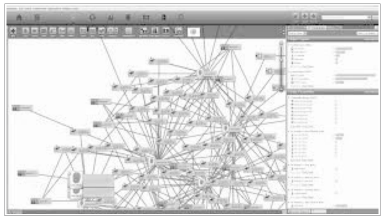
- **Computational Forensics** uses computational sciences to study **any type of evidence**:
  - Computer forensics
  - Crime Scene Investigation
  - Forensic palaeography
  - Forensic anthropology
  - Forensic chemistry
- **Computer Forensics** studies **digital evidence**:
  - File-system forensics
  - Live-system forensics
  - Mobile-device forensics etc.



Forensically-sound Computing Infrastructure

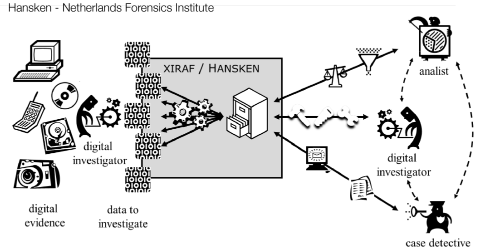


http://sqrrl.com/solutions/applications/cybersecurity/



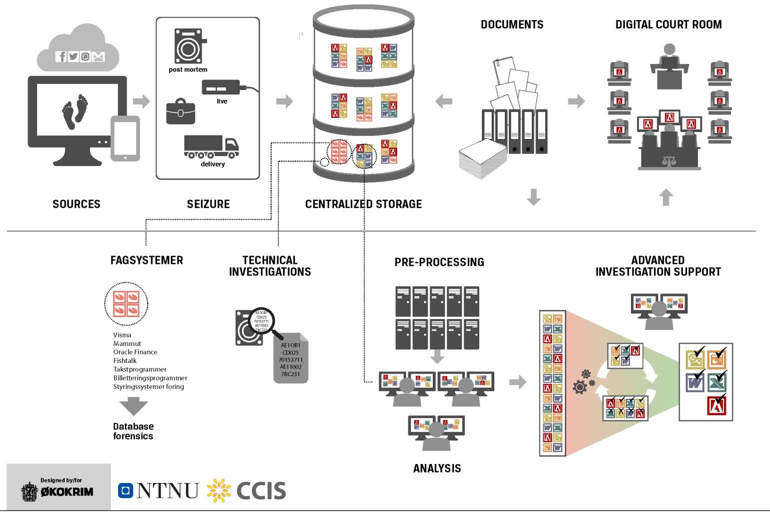
https://www.youtube.com/watch?v=6mlQmL2Lapw

Hansken - Netherlands Forensics Institute



Out of scope!
for this presentation

Project Example - Ars Forensica



25

Code-breaking Enigma, December 1942

26

Computing Machines & Intelligence (1950) by

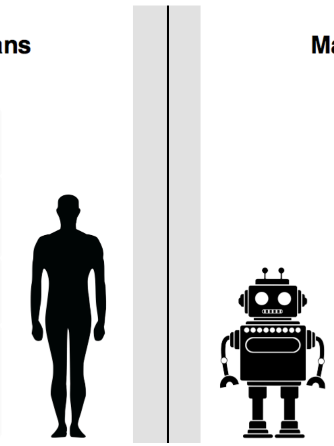
Alan Turing

https://wsimag.com/science-and-technology/36961-no-turing-test-for-consciousness 27

27

Hybrid-intelligence ?!

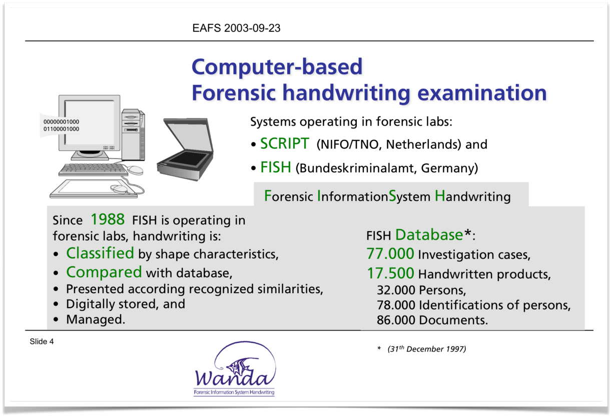
Humans	Machines
1. Computational Ability Humans are slow and likely to make mistakes	**1. Computational Ability** Machines are fast and near-flawless at computations
2. Random Number Generation Humans tend to 'spread out' number sequences.	**2. Random Number Generation** Machines less likely to 'spread out' numbers
3. Common Sense Humans have access to collective folk wisdom.	**3. Common Sense** Machines lack access to collective folk wisdom
4. Rationality Humans rely on biases and heuristics that deviate from the expectations of rational choice theory.	**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 28

28

FISH - Fighting Terrorism, Germany since 1975



EAFS 2003-09-23

Computer-based Forensic handwriting examination

Systems operating in forensic labs:
- SCRIPT (NIFO/TNO, Netherlands) and
- FISH (Bundeskriminalamt, Germany)

Forensic InformationSystem Handwriting

Since 1988 FISH is operating in forensic labs, handwriting is:
- Classified by shape characteristics,
- Compared with database,
- Presented according recognized similarities,
- Digitally stored, and
- Managed.

FISH Database*:
77.000 Investigation cases,
17.500 Handwritten products,
32.000 Persons,
78.000 Identifications of persons,
86.000 Documents.

Slide 4

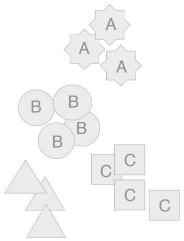
* (31th December 1997)

Fraunhofer IPK

Machine Learning & Pattern Recognition

Fundamentals

Machine Learning & Pattern Recognition

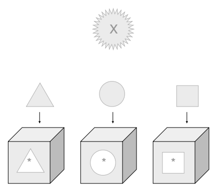


Pattern
- "as opposite of a chaos; it is an entity, vaguely defined, that could be given a name" Watanabe 1985

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

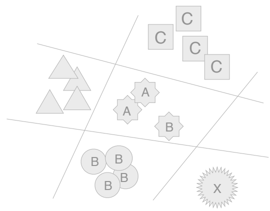
Same Facet - Different Origin
- Machine Learning - Computer Science
- Patter Recognition / Data Mining - Engineering
- Predictive Analytics - Business / Marketing

Pattern Classification



Supervised Classification
pre-defined by the
system designer

Machine Learning

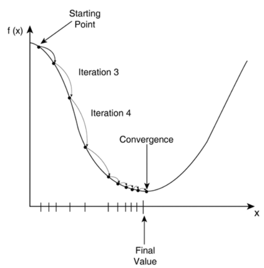


Unsupervised Classification
learning based on the
similarity of pattern

Data Mining

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).



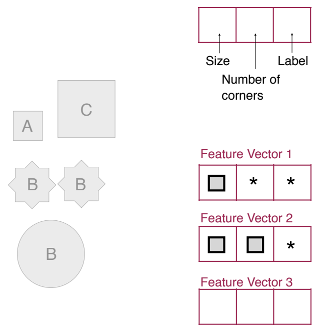
Representation of Pattern Characteristics

Goal

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

Tasks

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



Pattern Representation & Classification

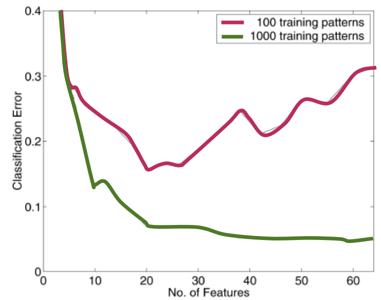
	x	A	C	A	B	B	
Feature Vector 1: □ * *		1 * *	2 * *	1 * *	1 * *	2 * *	2 (2)
Feature Vector 2: □ □ *		1 4 *	2 4 *	1 6 *	1 6 *	2 0 *	4 (6)
Feature Vector 3: □ □ □		1 4 A	2 4 C	1 6 A	1 6 B	2 0 B	5 (18)

Size | Number of corners | Label

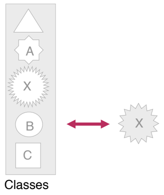
Classes

Classifier Training, … How do Computers learn?

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



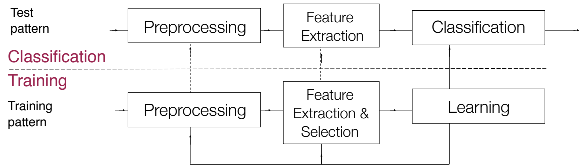
Classification & Matching



- Identification 1:N comparison
- To which class is the pattern assigned ?

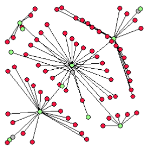
- Verification 1:1 comparison
- Are the reference and the pattern similar ?

Model for Pattern Classification

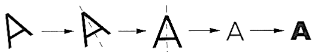


Statistical Pattern Recognition: A Review, A.K. Jain, R.P.W. Duin and J. Mao, 2000, PAMI
Note that biological-inspired methods come in addition

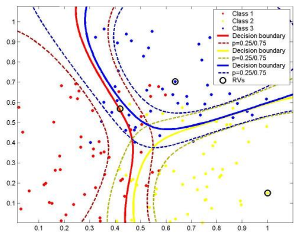
Commonly known Pattern-Recognition Approaches



- Template Matching

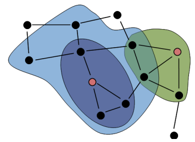


- Syntactical or Structural PR
- Statistical PR
- Neural Networks



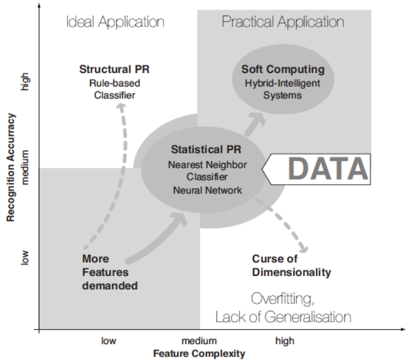
Statistical PR in Numbers

- 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

Towards Data-driven Approaches



BIG DATA Analytics

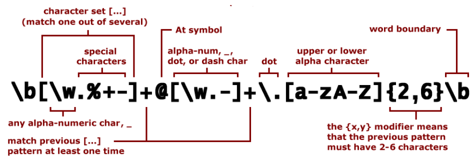
Inter-relation of
feature complexity and
expected recognition accuracy.

Reference: Franke (2005)

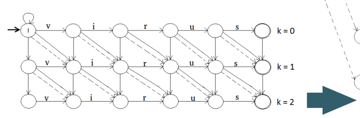
41

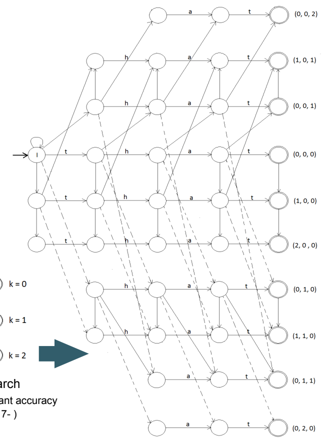
41

Regular Expressions vs. Approximate String Matching



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-)

42

42

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.

43

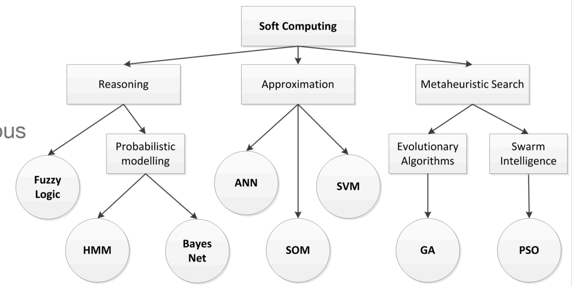
43

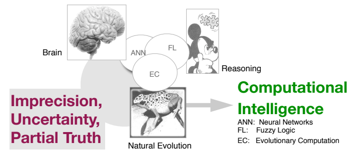
Data Science | Machine Learning & Computational Intelligence

44

Requirements on Computational Methods

Large scale Forensic Investigations
- Situation-aware methods
- Quantified, measurable indicators
- Adaptive, self-organising models
- Distributed, cooperative, autonomous



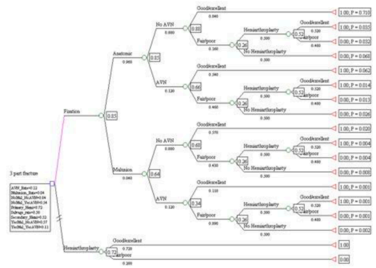


ANN: Neural Networks
FL: Fuzzy Logic
EC: Evolutionary Computation

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

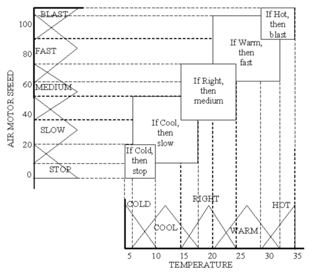
Hard Computing vs. Soft Computing

Decision Tree



www.data-machine.com

Fuzzy Rules



orthojournal.wordpress.com

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

Computational Forensics | Scientific Computing in Forensics

Admission of Computational Forensics

- **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

- **Education and training**, Revealing the state-of-the art in *each* domain
- **Sources of information** on events, activities and financing opportunities
- **International forum to peer-review and exchange**, e.g., IWCF workshops
- **Performance evaluation, benchmarking, proof and standardization** of algorithms
- Resources in forms of **data sets, software tools, and specifications** e.g. data formats
- **New Insights** on problem description and procedures

- Questions on methods for **dimensionality reduction** - loss of relevant information
- Questions on **extracted numerical parameters** - loss of information due to inappropriate features
- Questions on the reliability of **applied computational method / tool**
- Questions on the final conclusion due to **“wrong” computational results**

- Computational forensics holds the potential to greatly benefit all of the forensic sciences.
- For the computer scientist it poses a new frontier where new problems and challenges are to be faced.
- The potential benefits to society, meaningful inter-disciplinary research, and challenging problems should attract high quality students and researchers to the field.

49

49

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

– John Dewel

50

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
Email: katrin.franke@ntnu.no | kyfranke@ieee.org
Skype: kyfranke | www.kyfranke.com

51

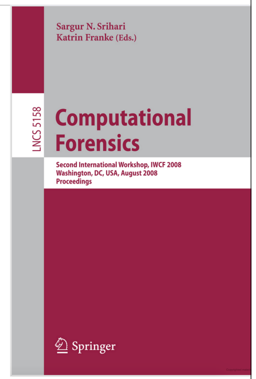
Katrin Franke

- (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
- Academic Advisor to EUROPOL, European Cybercrime Center (EC3), 2014-present
- Academic 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
  - Internet 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

52

52