Lessons from the Past, Challenges for the Future
(The EC09 Evaluation Framework in the Deep Learning Era)

François-Xavier Standaert, Tal Malkin, Moti Yung

UCLouvain, Columbia University, Google

Distinguished Lectures on Security & Privacy
IIT Kharagpur, May 3, 2022

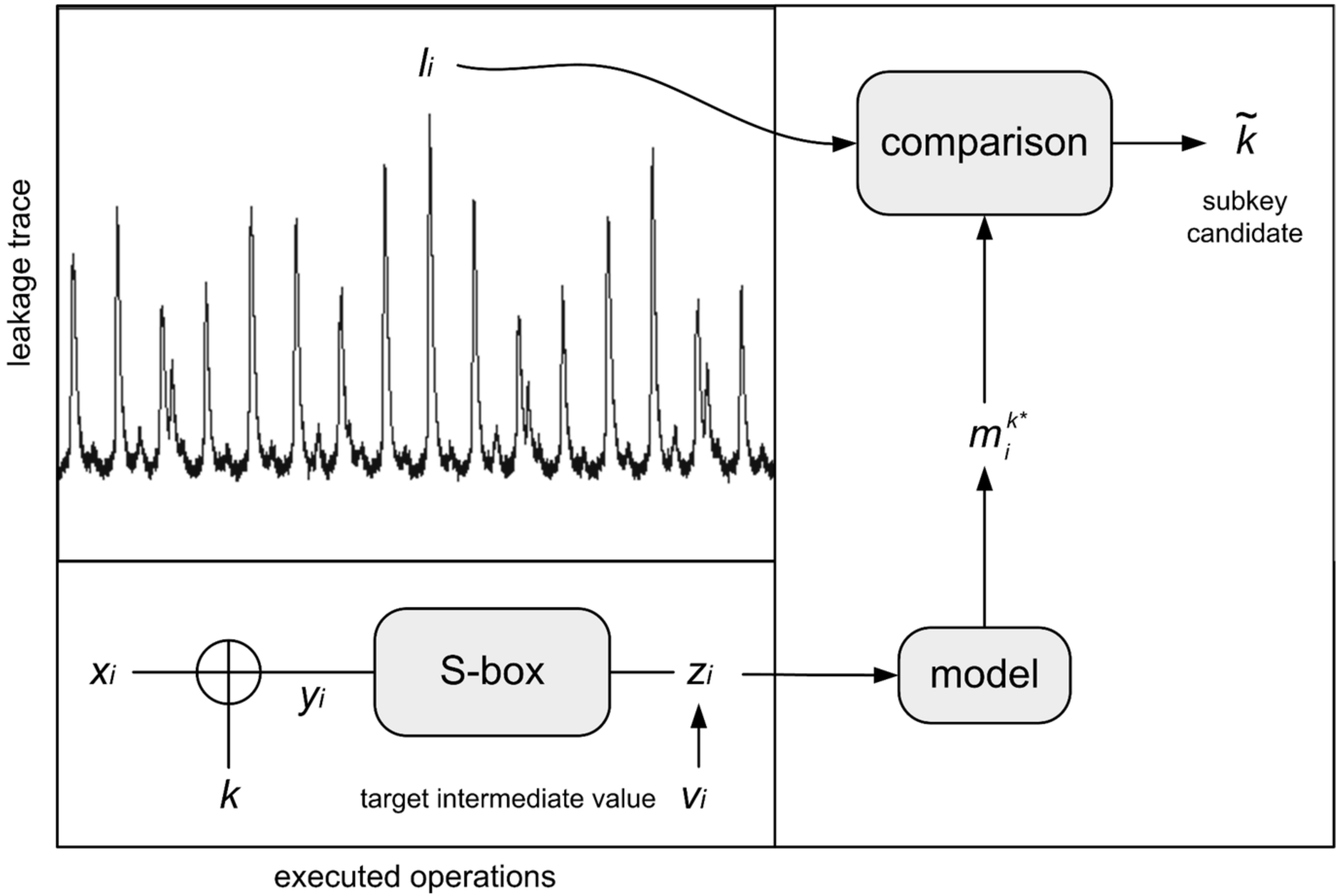
Outline

- The situation 15 years ago
- The EC09 evaluation framework
- Challenges and (partial) solutions
- Deep learning: what is new?
- Conclusions (technical & non-technical)

Outline

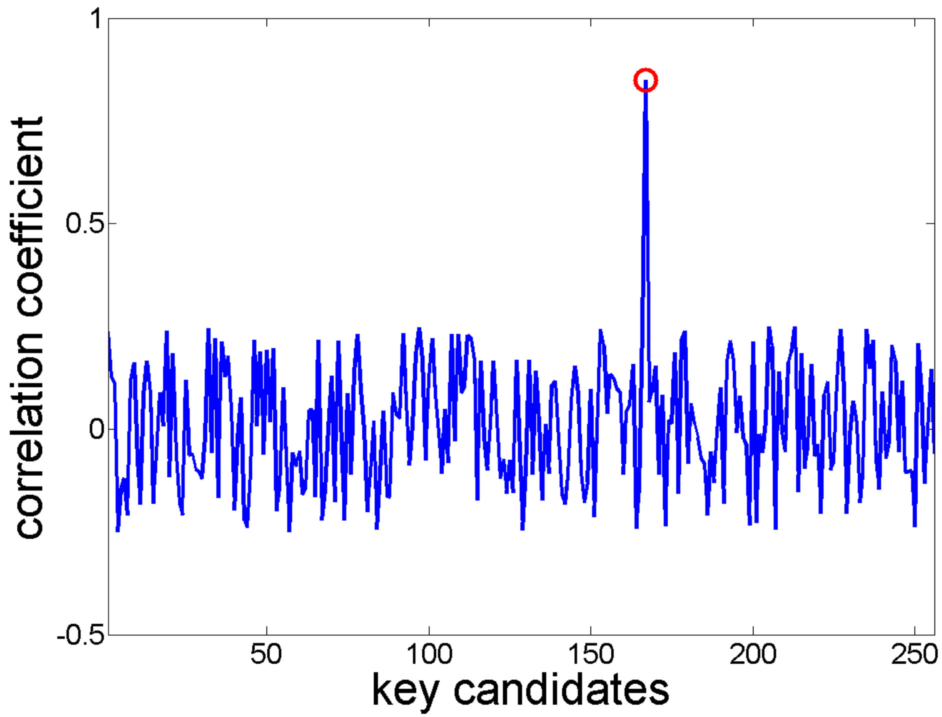
- **The situation 15 years ago**
- The EC09 evaluation framework
- Challenges and (partial) solutions
- Deep learning: what is new?
- Conclusions (technical & non-technical)

Power analysis attacks [KJJ99]



How to evaluate these attack (I) ?

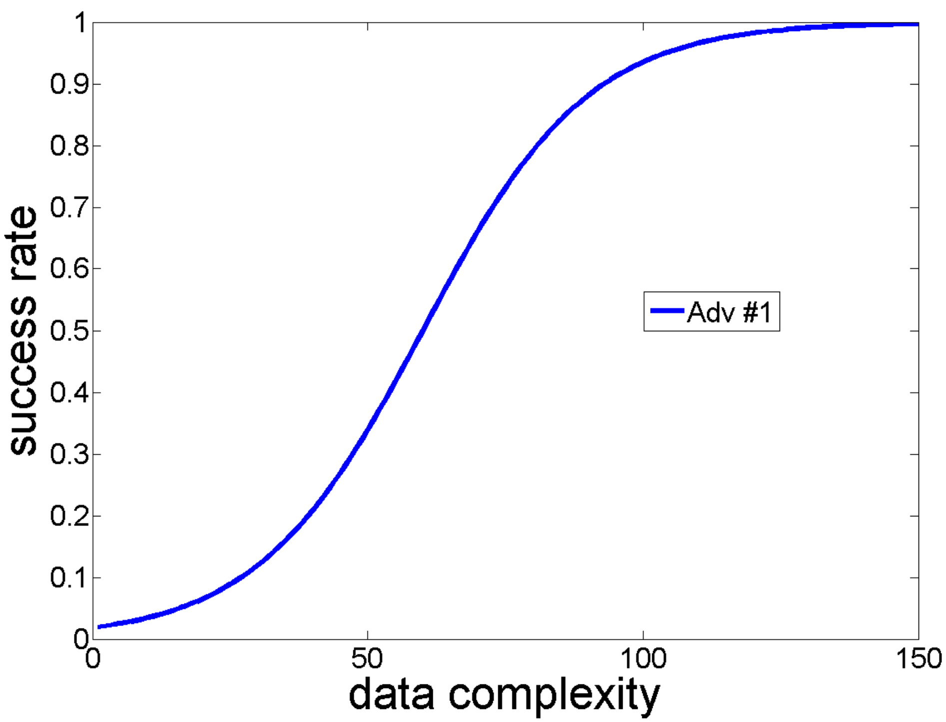
- Launch an attack with an arbitrary distinguisher



- But is gives no statistical confidence

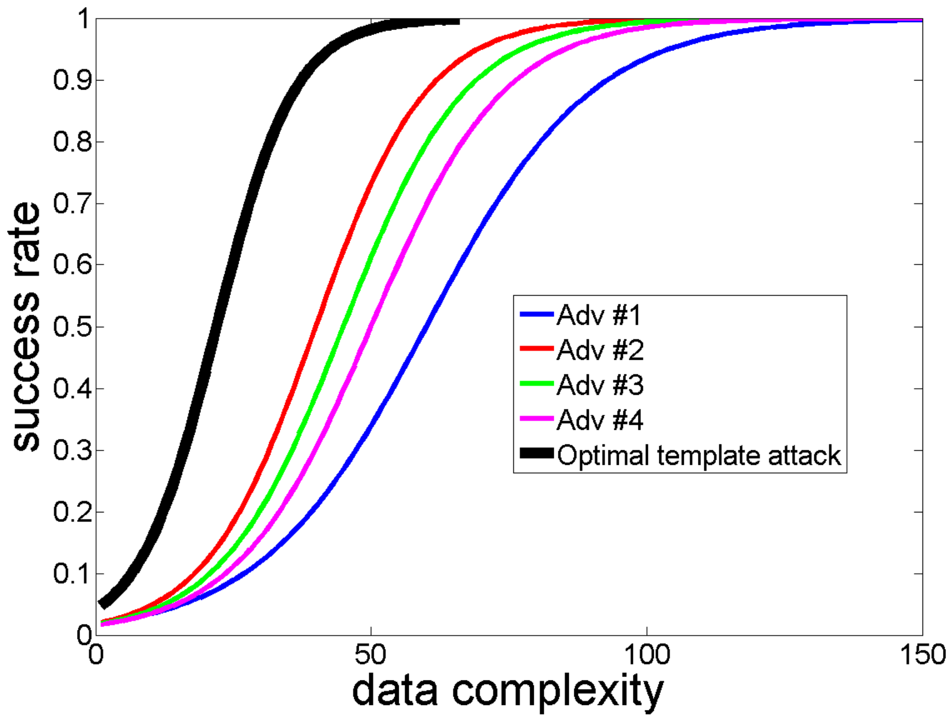
How to evaluate these attack (II) ?

- Repeat the attack and estimate a success rate



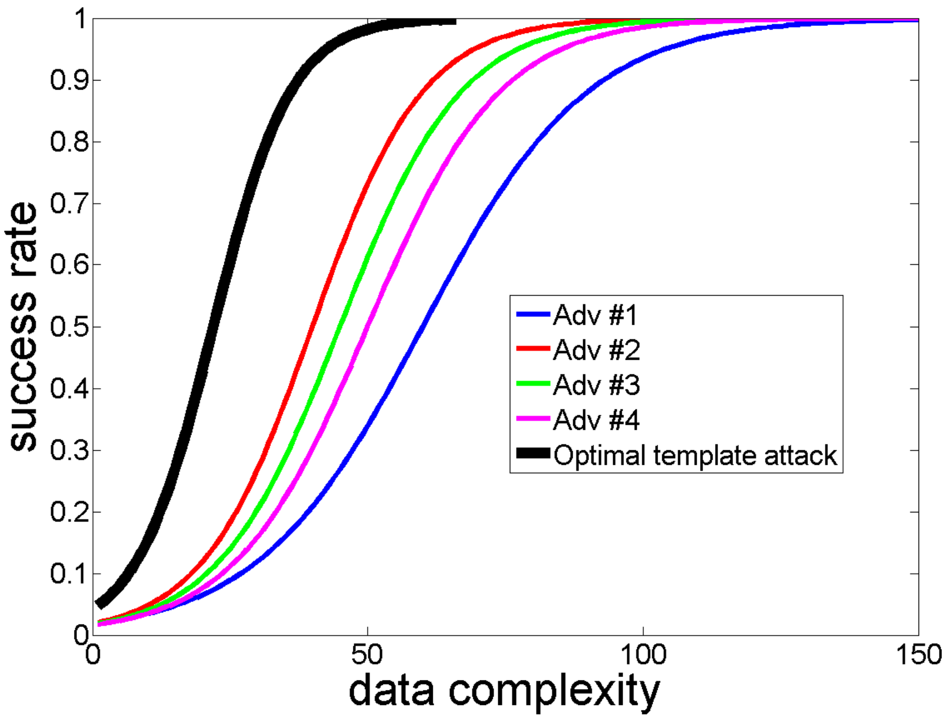
- But the adversary can still be subpotimal

- Try to find out what is the « optimal » attack?



How to evaluate these attack (III) ?

- Try to find out what is the « optimal » attack?



- Or to find out whether it is « practical »?

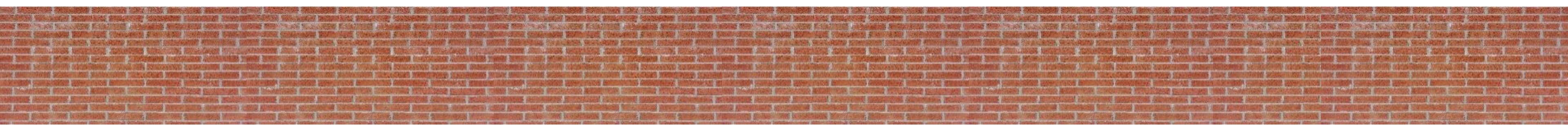
How to evaluate these attack (IV) ?

- Try to find out what is the « optimal » attack?

⇒ Worst-case academic (cryptographic) approach

≈ Kerckhoffs’ laws at the implementation level

  - Goal: formalize and develop long-term security



  - Goal: fix an emergency situation efficiently

≈ Rate attacks based on « adversary’s potential »

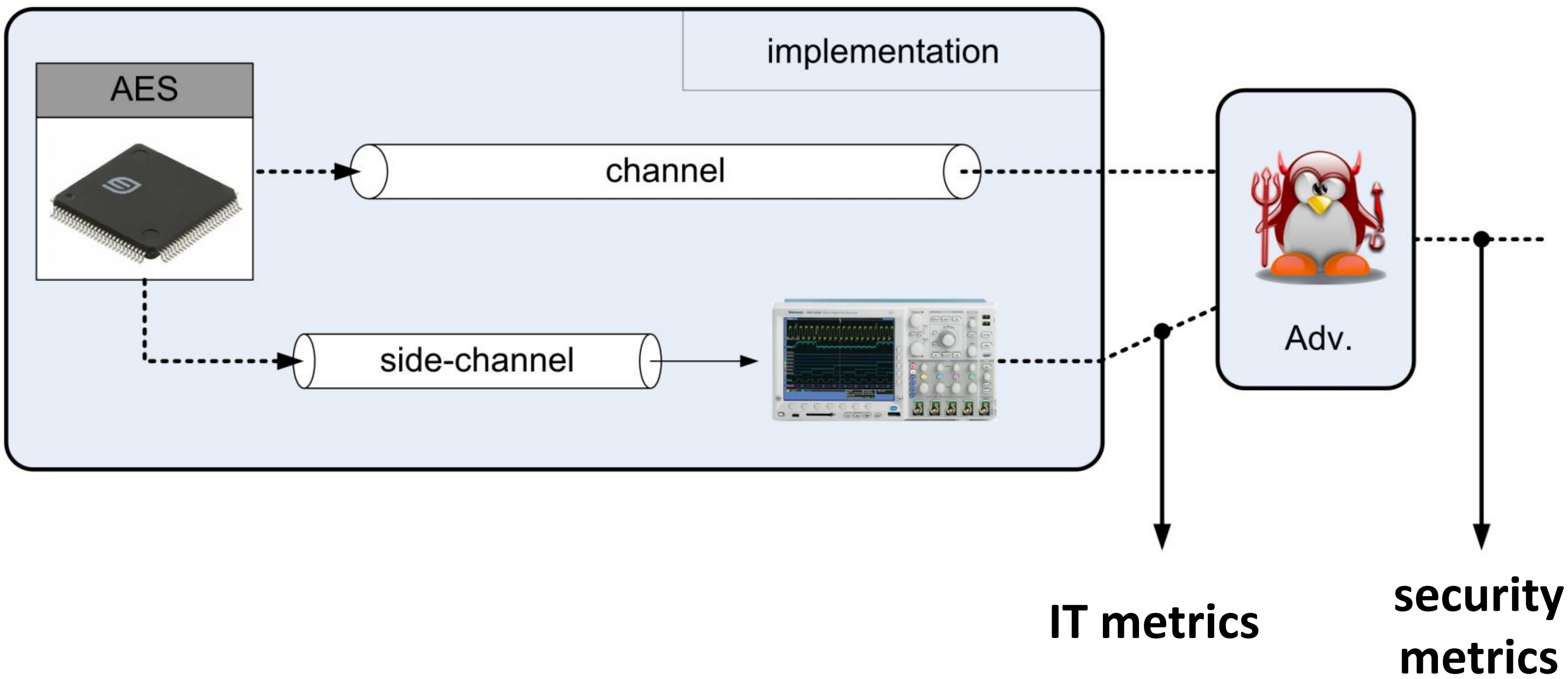
⇒ Industrial evaluation/certification schemes

- Or to find out whether it is « practical »?

Outline

- The situation 15 years ago
- **The EC09 evaluation framework**
- Challenges and (partial) solutions
- Deep learning: what is new?
- Conclusions (technical & non-technical)

Formalizing worst-case attacks [SMY09]



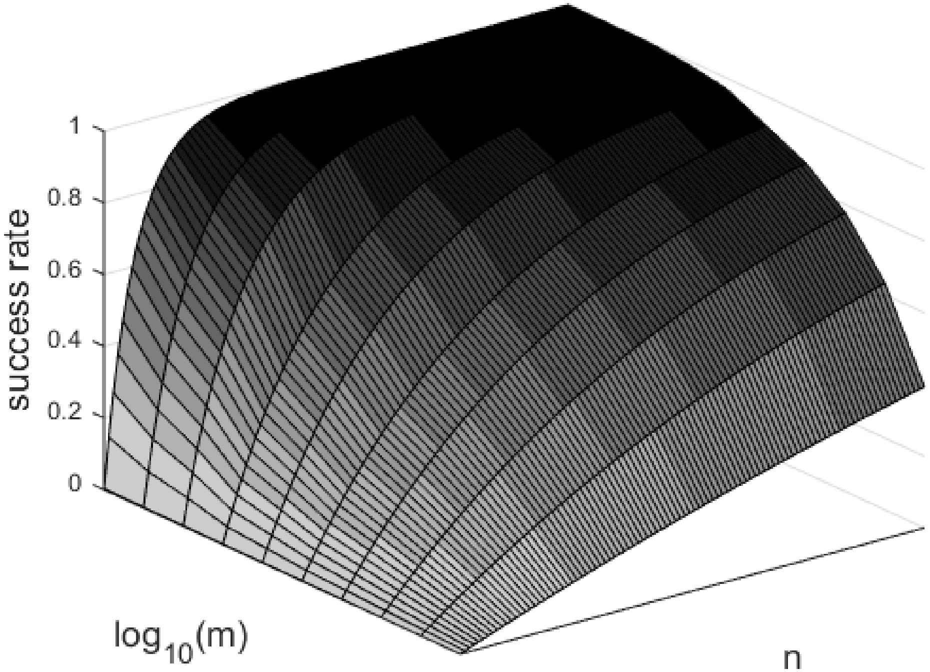
- Conceptual separation between metrics
  - IT metrics (e.g., MI,PI) $\perp\!\!\!\perp$ of adv.'s comp. power
  - Security metrics (e.g., SR, GE) $\propto$ adv.'s comp. power

Why two metrics (I)?

- Security metrics are more « complete »

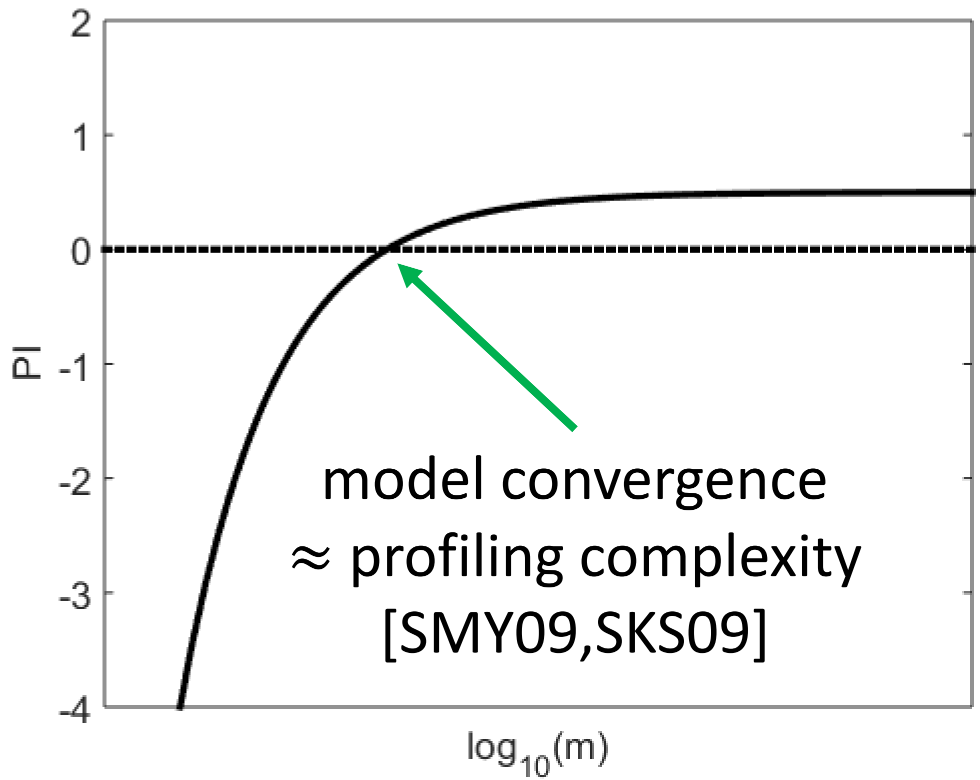
⇒ why not directly going for worst-case SR or GE?

- Problem: can be quite expensive to estimate
  - E.g., m traces to train model & n traces to attack



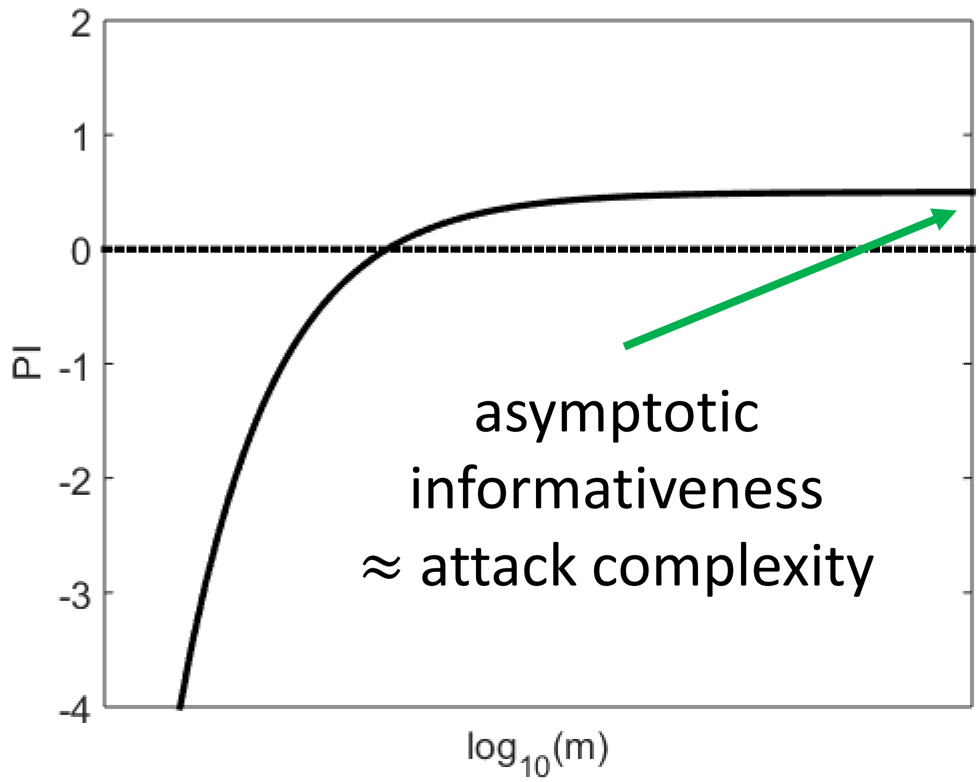
Why two metrics (II)?

- IT metrics enable more efficient evaluations
  - That are easier to interpret (≈ learning curves)



Why two metrics (II)?

- IT metrics enable more efficient evaluations
  - That are also easier to interpret visually



$$\text{n(SR=90\%)} \approx \frac{cst}{\mathrm{PI}(K;L)}$$

- [SMY09,MOS11]: specific leakages
- Later generalized in [DFS15,dCGRP19]

Metrics intuitions

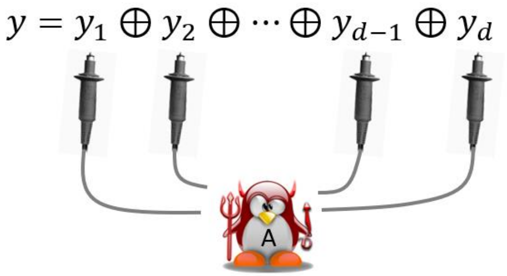
- When the attack complexity is fixed by design (i.e., in a **SPA** setting), use security metrics
- When the attack complexity is unknown (i.e., in a **DPA** setting) IT metrics provide a shortcut

Metrics intuitions

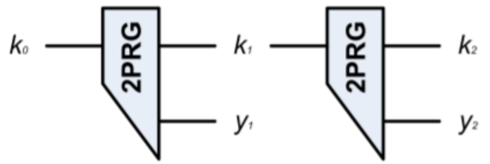
- When the attack complexity is fixed by design (i.e., in a **SPA** setting), use security metrics
- When the attack complexity is unknown (i.e., in a **DPA** setting) IT metrics provide a shortcut

⇒ Framework ≈ middleware btw. models & devices

- IT metrics ≈ noise assumption needed in masking proofs [PR13]



- Security metrics ≈ leakage bound in leak-resilience proofs [DP08]

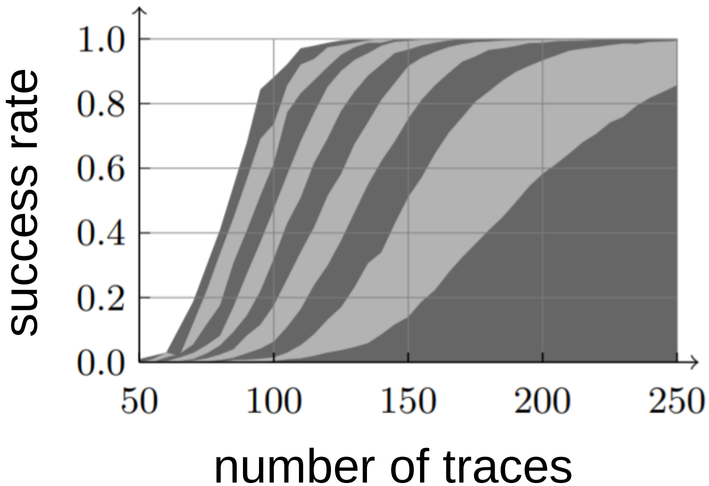


Outline

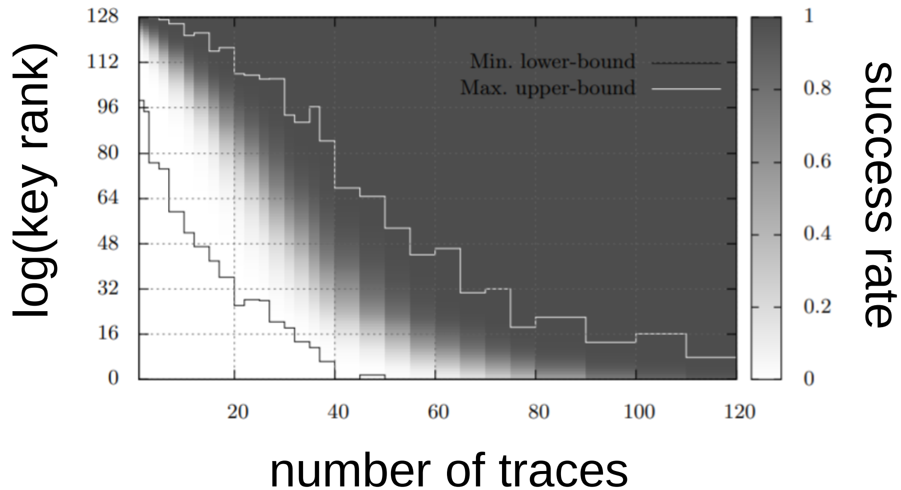
- The situation 15 years ago
- The EC09 evaluation framework
- **Challenges and (partial) solutions**
- Deep learning: what is new?
- Conclusions (technical & non-technical)

(Nearly) solved challenge: security metrics

- SR & GE easy to estimate for 8-bit subkeys
  - How to do it for full (e.g., 128-bit) keys?



key enumeration



rank estimation

- Reasonably well solved [VGRS12,VGS13]
  - Many follow ups work and optimizations

Unsolved Challenge: IT metrics (I)

$$\mathrm{MI}(K;L) = \mathrm{H}[K] + \sum_{k\in K} \Pr[k] \cdot \int \mathrm{f}_{\mathrm{real}}(l|k) \cdot \log_2\left(\Pr_{\mathrm{real}}[k|l]\right)$$

- With $\Pr_{\mathrm{real}} = \frac{\mathrm{f}_{\mathrm{real}}(l|k)}{\sum_{k^*} \mathrm{f}_{\mathrm{real}}(l|k^*)}$ and $\mathrm{f}_{\mathrm{real}}(l|k)$ unknown!

Unsolved Challenge: IT metrics (I)

$$\mathrm{MI}(K;L) = \mathrm{H}[K] + \sum_{k \in K} \Pr[k] \cdot \int \mathrm{f}_{\mathrm{real}}(l|k) \cdot \log_2\left(\Pr_{\mathrm{real}}[k|l]\right)$$

- With $\Pr_{\mathrm{real}} = \frac{\mathrm{f}_{\mathrm{real}}(l|k)}{\sum_{k^*} \mathrm{f}_{\mathrm{real}}(l|k^*)}$ and $\mathrm{f}_{\mathrm{real}}(l|k)$ unknown!
- Information that can be extracted with a model

$$\mathrm{PI}(K;L) = \mathrm{H}[K] + \sum_{k \in K} \Pr[k] \cdot \int \mathrm{f}_{\mathrm{real}}(l|k) \cdot \log_2\left(\Pr_{\mathrm{model}}[k|l]\right)$$

$$\mathrm{MI}(K;L) = \mathrm{H}[K] + \sum_{k\in K} \Pr[k] \cdot \int \mathrm{f}_{\mathrm{real}}(l|k) \cdot \log_2\left(\Pr_{\mathrm{real}}[k|l]\right)$$

- With $\Pr_{\mathrm{real}} = \frac{\mathrm{f}_{\mathrm{real}}(l|k)}{\sum_{k^*} \mathrm{f}_{\mathrm{real}}(l|k^*)}$ and $\mathrm{f}_{\mathrm{real}}(l|k)$ unknown!
- Information that can be extracted with a model

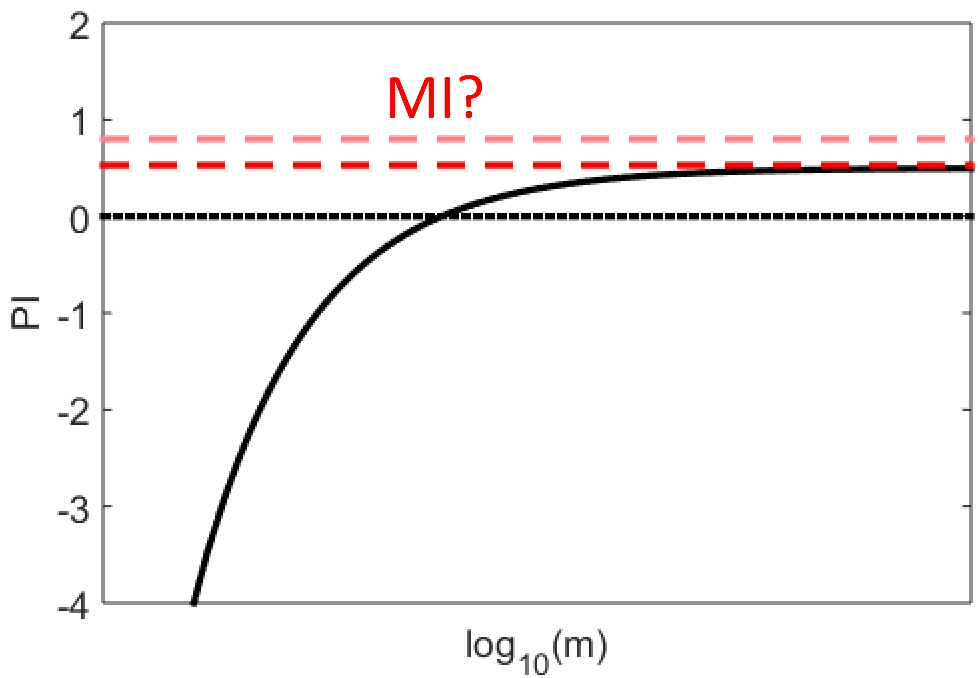
$$\mathrm{PI}(K;L) = \mathrm{H}[K] + \sum_{k\in K} \Pr[k] \cdot \int \mathrm{f}_{\mathrm{real}}(l|k) \cdot \log_2\left(\Pr_{\mathrm{model}}[k|l]\right)$$

- Which can be evaluated by sampling in 2 steps

$$\widehat{\mathrm{PI}}(K;L) = \mathrm{H}[K] + \sum_{k\in K} \Pr[k] \cdot \sum_{l' \overset{N_t}{\leftarrow} \mathrm{f}_{\mathrm{real}}(l|k)} \frac{1}{N_t} \cdot \log_2\left(\widehat{\Pr}_{\mathrm{model}}[k|l']\right)$$

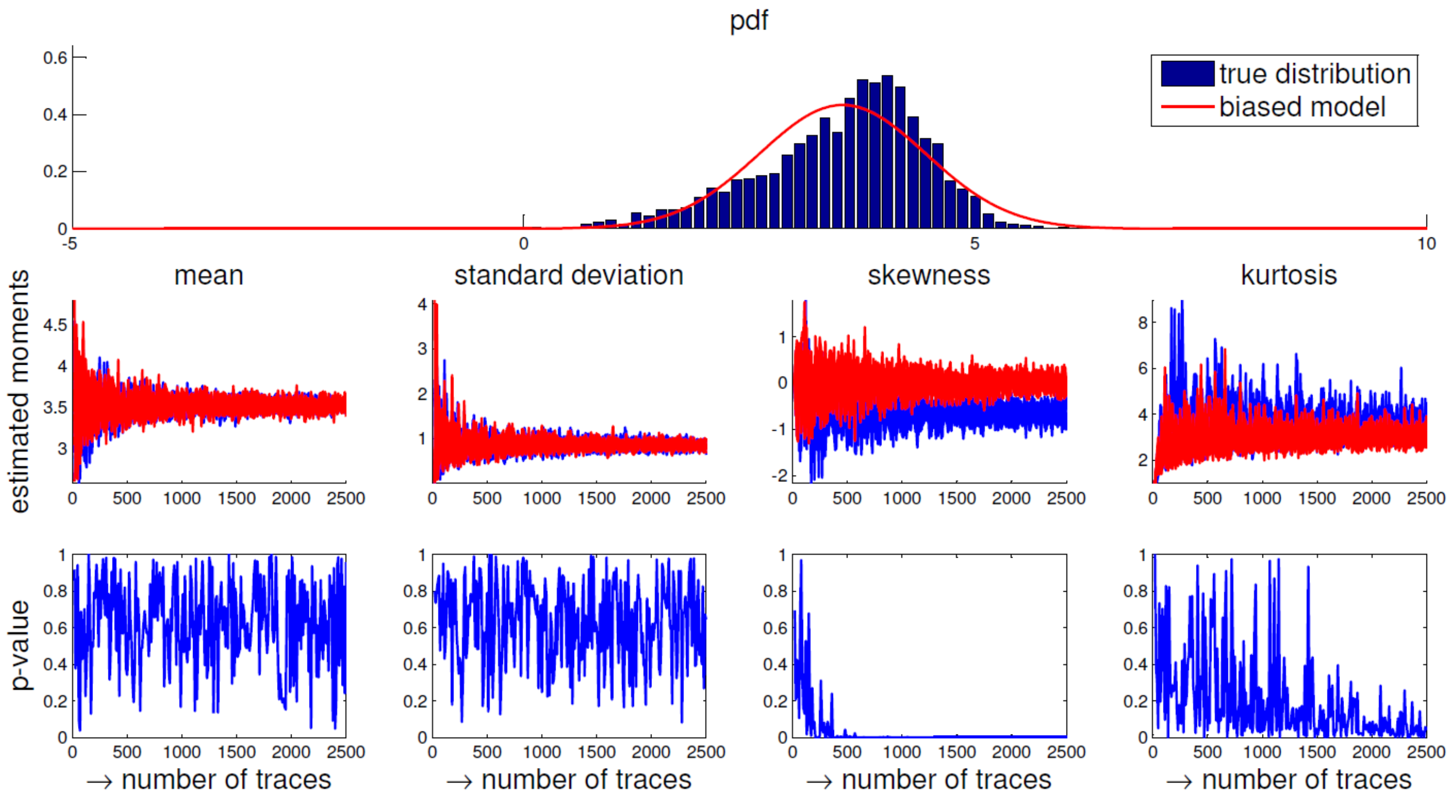
Unsolved Challenge: IT metrics (II)

- Worst-case (MI=PI) never happens in practice
  - Requires a perfect knowledge of the leakage model
- Evaluator question: how large is the gap?



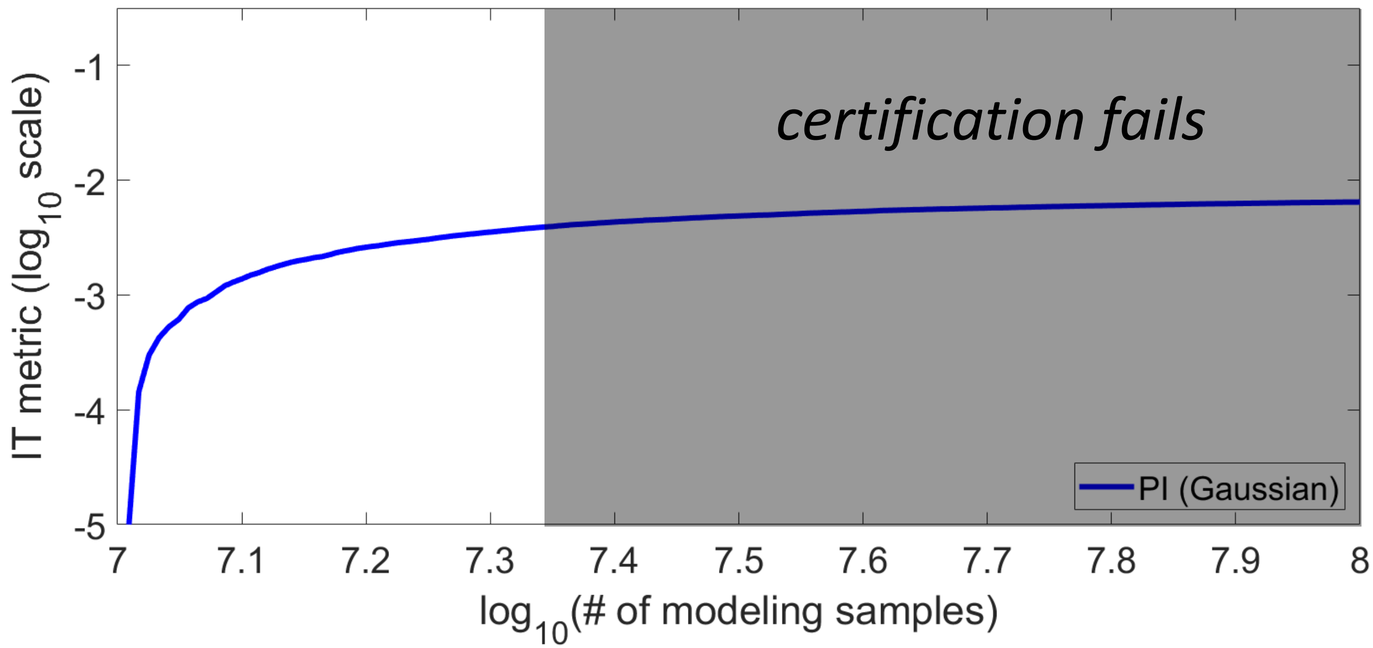
Unsolved Challenge: IT metrics (III)

- Qualitative attempt [DSV14]: model good enough if assumption errors small << estimation errors



Unsolved Challenge: IT metrics (III)

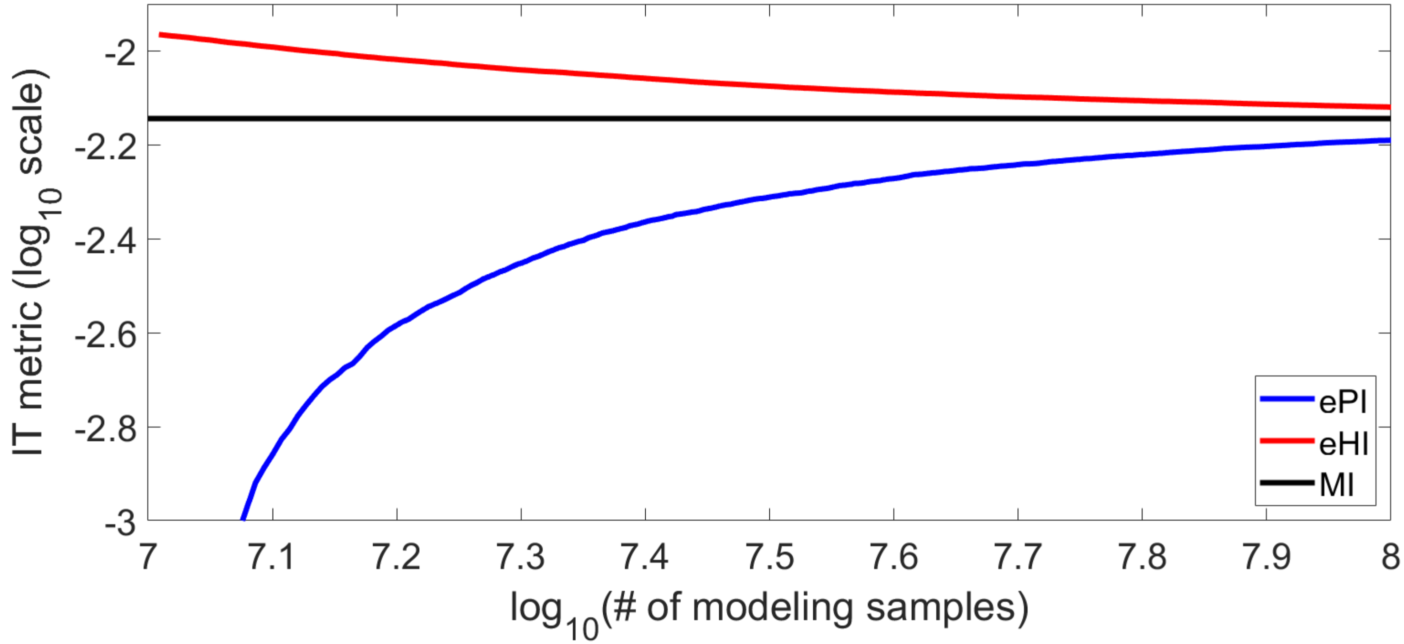
- Qualitative attempt [DSV14]: model good enough if assumption errors small << estimation errors



- Does not say anything about the size of the gap

Unsolved Challenge: IT metrics (IV)

- Quantitative attempt [B+19]: upper bound the MI thanks to the HI (≈ training information)



- Limited to models based on the empirical distrib.
  - *Open problem: high-order & multivariate leakages*

Outline

- The situation 15 years ago
- The EC09 evaluation framework
- Challenges and (partial) solutions
- **Deep learning: what is new?**
- Conclusions (technical & non-technical)

Do we need new metrics?

- Revisiting evaluation metrics [P+19]

The Curse of Class Imbalance and Conflicting Metrics with Machine Learning for Side-channel Evaluations

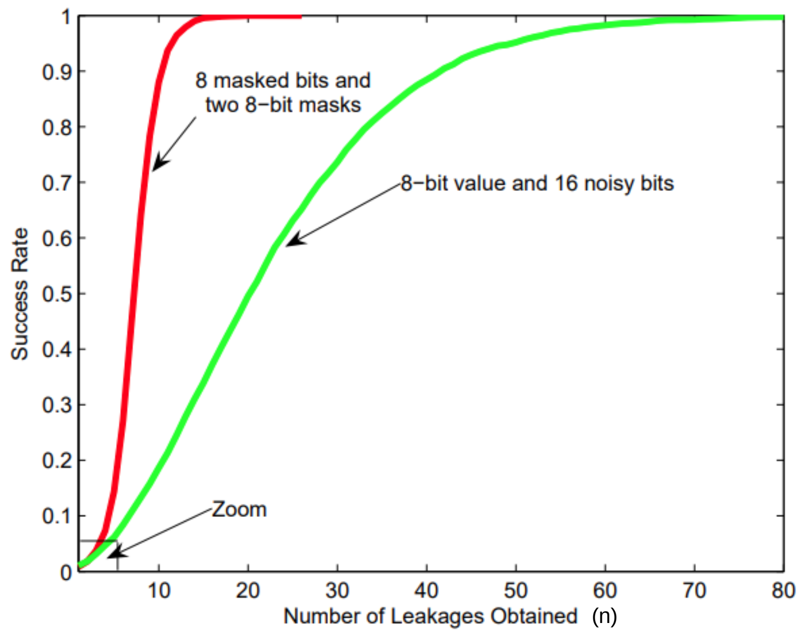
Stjepan Picek[1,2], Annelie Heuser[3], Alan Jovic[4], Shivam Bhasin[5] and Francesco Regazzoni[6]

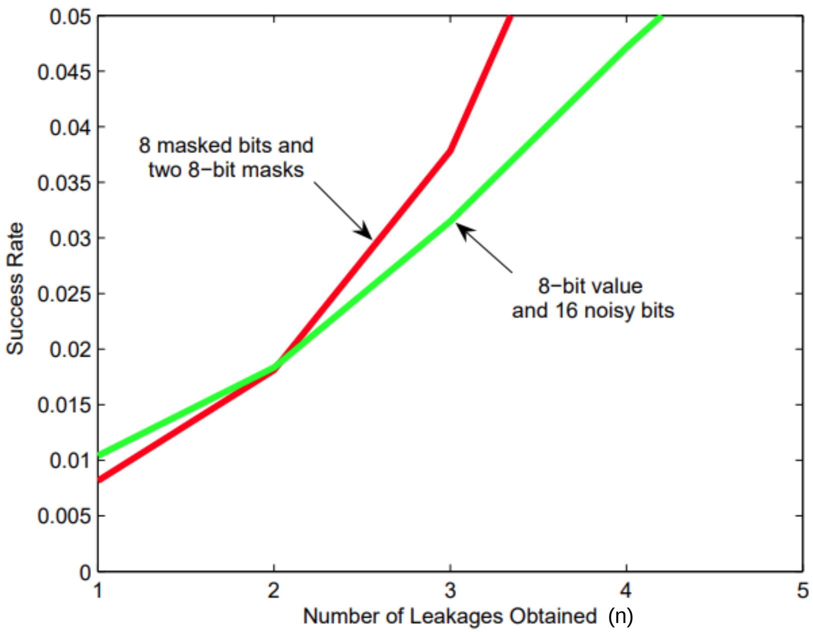
[1] Delft University of Technology, Delft, The Netherlands
s.picek@tudelft.nl
[2] LAGA, Department of Mathematics, University of Paris 8 (and Paris 13 and CNRS), France
[3] Univ Rennes, Inria, CNRS, IRISA, France
annelie.heuser@irisa.fr
[4] University of Zagreb Faculty of Electrical Engineering and Computing, Croatia
alan.jovic@fer.hr
[5] Physical Analysis and Cryptographic Engineering, Temasek Laboratories at Nanyang Technological University, Singapore
sbhasin@ntu.edu.sg
[6] University of Lugano, Switzerland
regazzoni@alari.ch

Abstract. We concentrate on machine learning techniques used for profiled side-channel analysis in the presence of imbalanced data. Such scenarios are realistic and often occurring, for instance in the Hamming weight or Hamming distance leakage models. In order to deal with the imbalanced data, we use various balancing techniques and we show that most of them help in mounting successful attacks when the data is highly imbalanced. Especially, the results with the SMOTE technique are encouraging, since we observe some scenarios where it reduces the number of necessary measurements more than 8 times. Next, we provide extensive results on comparison of machine learning and side-channel metrics, where we show that machine learning metrics (and especially accuracy as the most often used one) can be extremely deceptive. This finding opens a need to revisit the previous works and their results in order to properly assess the performance of machine learning in side-channel analysis.

- ∃? a metric issue specific to machine learning
  - Or is it related to the context (SPA vs. DPA)?

Back 15 years ago [SPAQ06]





- SR(n) can be a bad predictor of DPA complexity
  - Because n needed for high SR is not known a priori
    - Which motivated the introduction of IT metrics

Do we need new metrics?

- ∃? a metric issue specific to machine learning
  - Or is it related to the context (SPA vs. DPA)?
- *Tentative answer*: it is true that accuracy (a security metric) can be deceptive in SCA evaluations, but the reason is the context (SPA or DPA), no the type of statistical learning tool

Do we need new metrics?

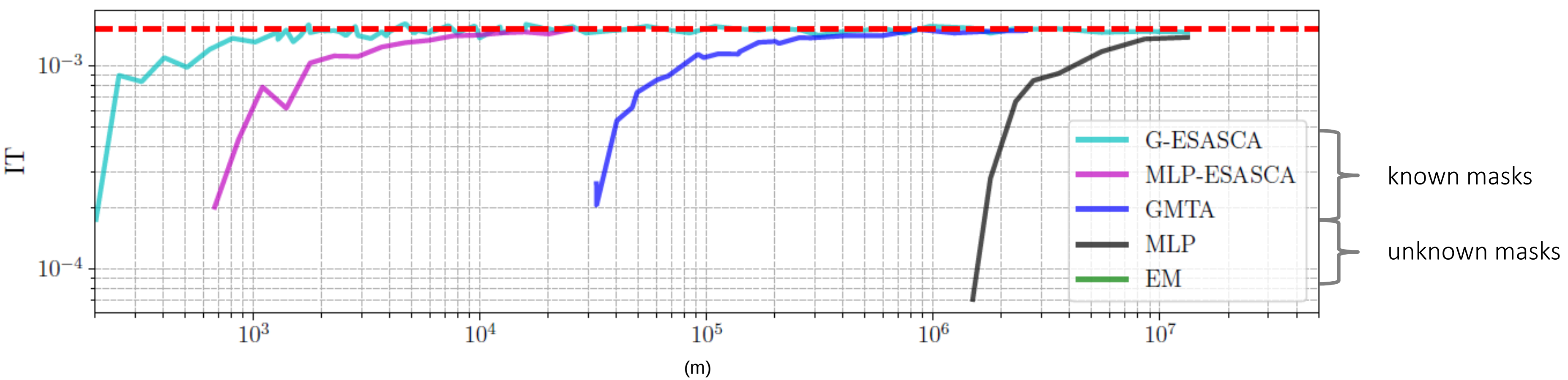
- ∃? a metric issue specific to machine learning
  - Or is it related to the context (SPA vs. DPA)?
- *Tentative answer*: it is true that accuracy (a security metric) can be deceptive in SCA evaluations, but the reason is the context (SPA or DPA), no the type of statistical learning tool

⇒ *Lesson from the past*:

  - Use security metrics for SPA evaluations
  - Use IT metrics for efficient DPA evaluations
- Corollary: use IT metrics as loss functions [MDP20]

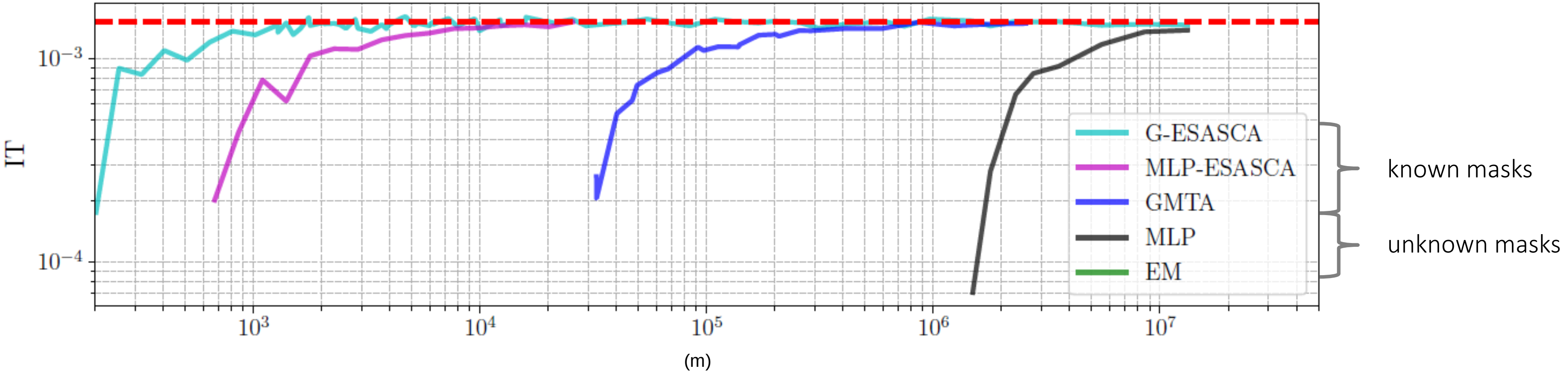
Bridging the gap

- Back to worst-case evaluations vs. practicality
  - Mostly differ in terms of adversary capabilities
    - E.g., implem. knowledge, profiling with known rand., …
- Machine learning can sometimes do in black box what worst-case attacks do with more capabilities [BDMS21]



Bridging the gap

- Back to worst-case evaluations vs. practicality
  - Mostly differ in terms of adversary capabilities
    - E.g., implem. knowledge, profiling with known rand., …
- Machine learning can sometimes do in black box what worst-case attacks do with more capabilities [BDMS21]

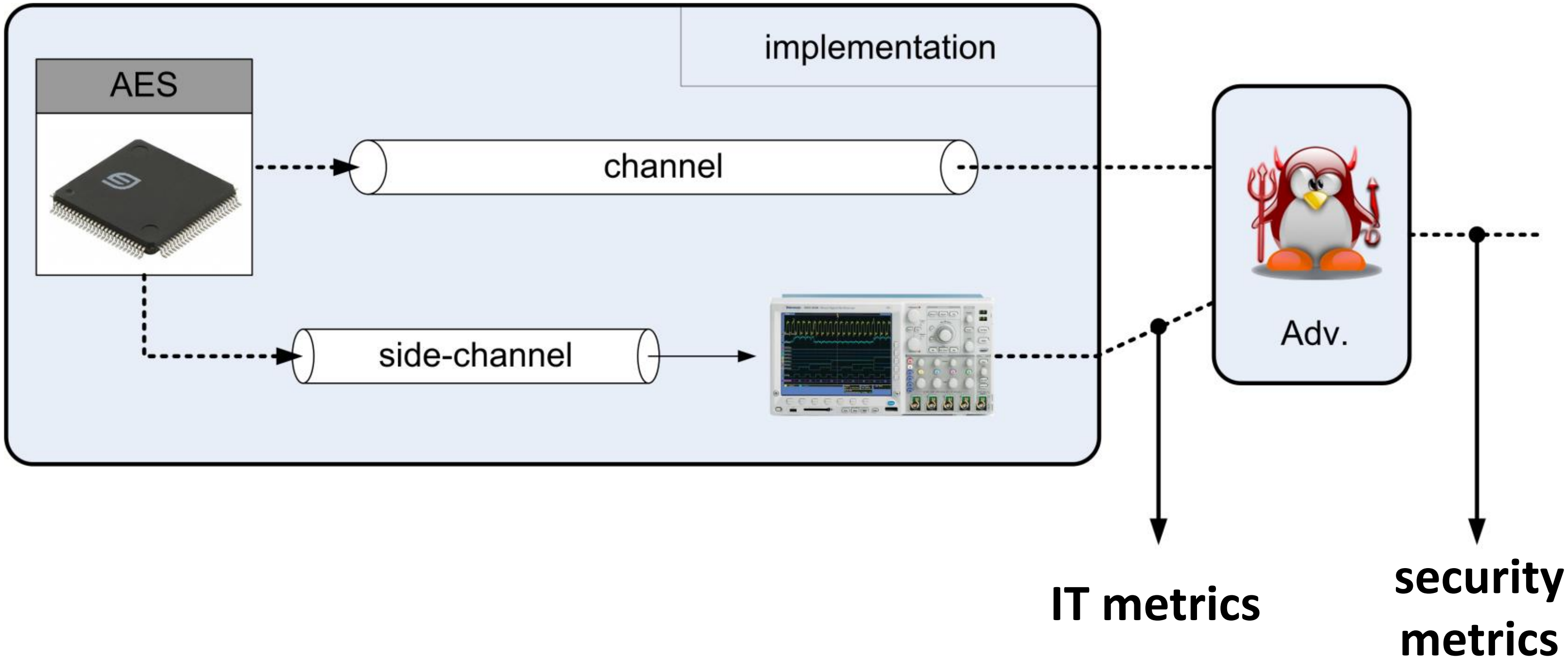


⇒ *Challenge for the future*: formalize this (lack of) gap

- New: deep learning (black box) convergence properties!

Outline

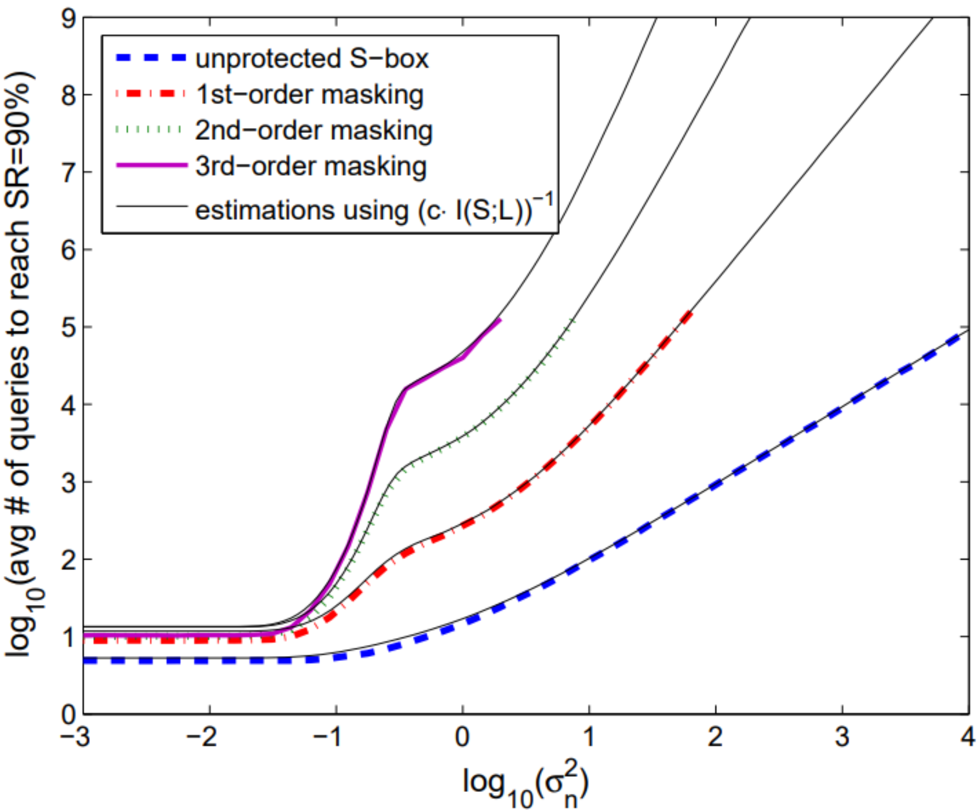
- The situation 15 years ago
- The EC09 evaluation framework
- Challenges and (partial) solutions
- Deep learning: what is new?
- **Conclusions (technical & non-technical)**



- Separating IT & security metrics is useful to
  - Structure evaluations (implem. vs. adv., SPA vs. DPA)
  - Serve as an interface with proofs (e.g., IT metrics for masking, security metrics for leakage-resilience)

Conclusion (present)

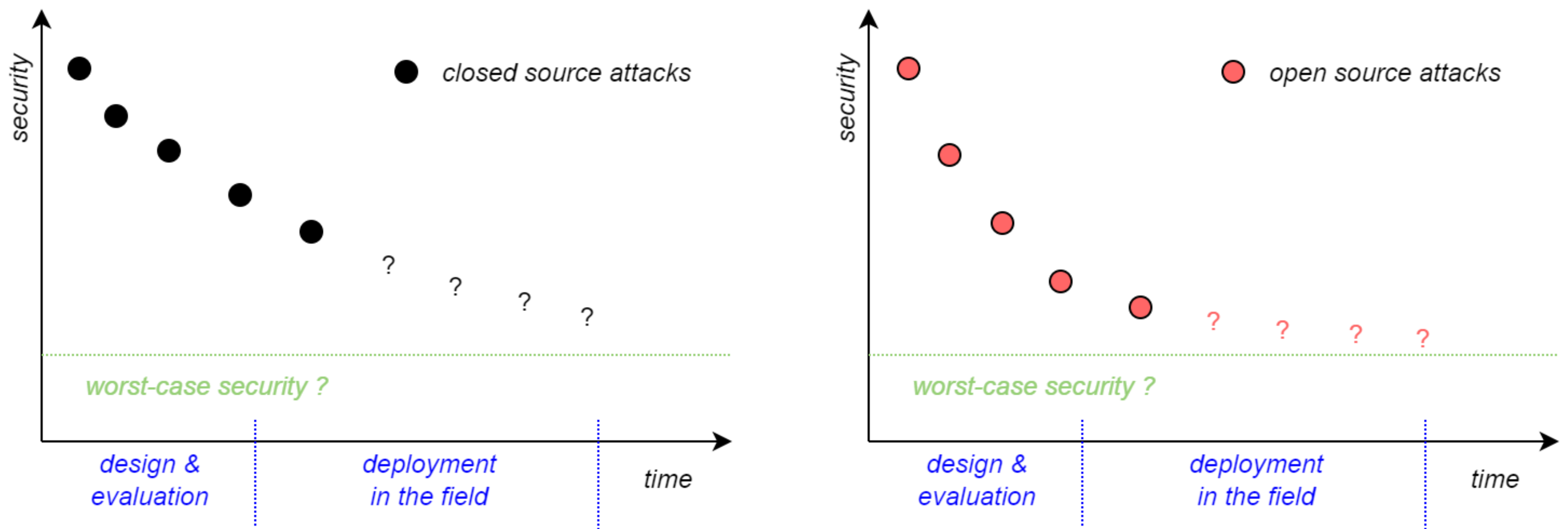
- IT metrics are a useful proxy before proving the security of a countermeasure & check tightness



- Masking [S+10]
  - Formally proven [PR13]
  - Still not tight [CFOS21]
- Shuffling [VMKS12]
  - Not proven yet
- Horizontal attacks [CS19]
  - Not proven yet
- Masking + shuffling [A+22]
- …

Conclusions (future)

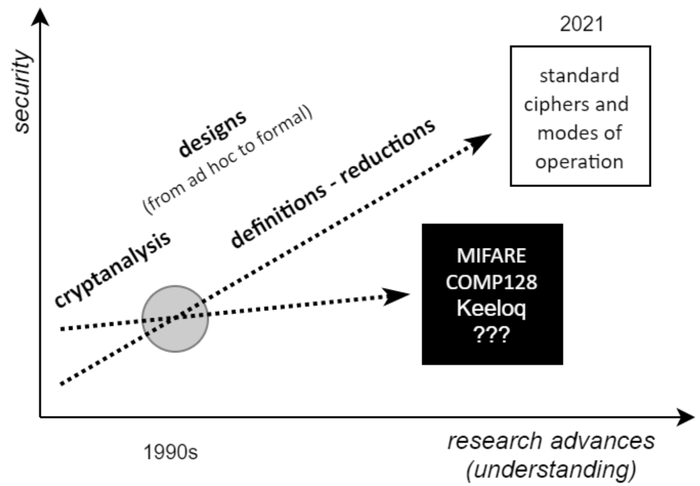
- Long-term security is hard to anticipate

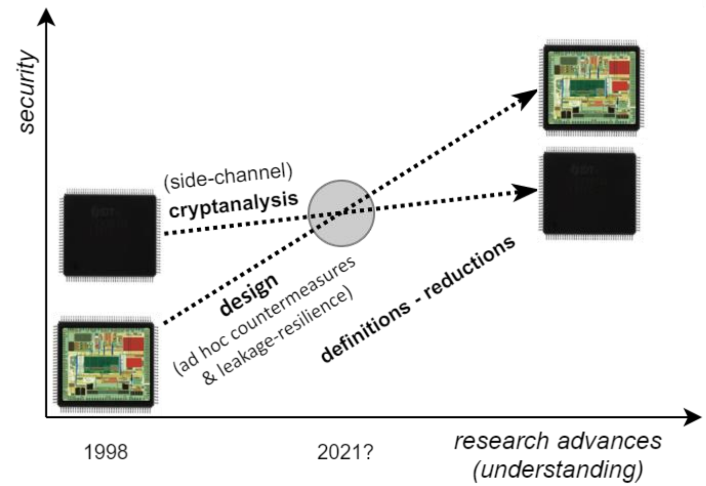


- Such anticipation is easier in an open setting
- Open problem: ∃? a gap btw. both appraches

Theory + practice (*rather than vs.*)

- For designs: push cryptographic formalism (≈ transparency) as far as possible ⇒ separate unambiguous assumptions from proofs





- Evaluations: start worst-case & then study relaxed adv. capabilities (i.e., backwards approach [A+20])

THANKS

http://perso.uclouvain.be/fstandae/

[KJJ99] Paul C. Kocher, Joshua Jaffe, Benjamin Jun: Differential Power Analysis. CRYPTO 1999: 388-397
[SMY09] François-Xavier Standaert, Tal Malkin, Moti Yung: A Unified Framework for the Analysis of Side-Channel Key Recovery Attacks. EUROCRYPT 2009: 443-461
[MOS11] Stefan Mangard, Elisabeth Oswald, François-Xavier Standaert: One for All - All for One: Unifying Standard DPA Attacks. IACR Cryptol. ePrint Arch. 2009: 449 (2009)
[DFS15] Alexandre Duc, Sebastian Faust, François-Xavier Standaert: Making Masking Security Proofs Concrete - Or How to Evaluate the Security of Any Leaking Device. EUROCRYPT (1) 2015: 401-429
[dCGRP19] Eloi de Chérisey, Sylvain Guilley, Olivier Rioul, Pablo Piantanida: Best Information is Most Successful Mutual Information and Success Rate in Side-Channel Analysis. IACR Trans. Cryptogr. Hardw. Embed. Syst. 2019(2): 49-79 (2019)
[PR13] Emmanuel Prouff, Matthieu Rivain: Masking against Side-Channel Attacks: A Formal Security Proof. EUROCRYPT 2013: 142-159
[DP08] Stefan Dziembowski, Krzysztof Pietrzak: Leakage-Resilient Cryptography. FOCS 2008: 293-302
[VGRS12] Nicolas Veyrat-Charvillon, Benoît Gérard, Mathieu Renauld, François-Xavier Standaert: An Optimal Key Enumeration Algorithm and Its Application to Side-Channel Attacks. Selected Areas in Cryptography 2012: 390-406
[VGS13] Nicolas Veyrat-Charvillon, Benoît Gérard, François-Xavier Standaert: Security Evaluations beyond Computing Power. EUROCRYPT 2013: 126-141
[DSV14] François Durvaux, François-Xavier Standaert, Nicolas Veyrat-Charvillon: How to Certify the Leakage of a Chip? EUROCRYPT 2014: 459-476
[B+19] Olivier Bronchain, Julien M. Hendrickx, Clément Massart, Alex Olshevsky, François-Xavier Standaert: Leakage Certification Revisited: Bounding Model Errors in Side-Channel Security Evaluations. CRYPTO (1) 2019: 713-737
[P+19] Stjepan Picek, Annelie Heuser, Alan Jovic, Shivam Bhasin, Francesco Regazzoni: The Curse of Class Imbalance and Conflicting Metrics with Machine Learning for Side-channel Evaluations. IACR Cryptol. ePrint Arch. 2018: 476 (2018)
[SPAQ06] François-Xavier Standaert, Eric Peeters, Cédric Archambeau, Jean-Jacques Quisquater: Towards Security Limits in Side-Channel Attacks. CHES 2006: 30-45
[MDP20] Loïc Masure, Cécile Dumas, Emmanuel Prouff: A Comprehensive Study of Deep Learning for Side-Channel Analysis. IACR Trans. Cryptogr. Hardw. Embed. Syst. 2020(1): 348-375 (2020)
[BDMS21] Olivier Bronchain, François Durvaux, François-Xavier Standaert, Loïc Masure,
Efficient Profiled Side-Channel Analysis of Masked Implementations, Extended. Preprint
[S+10] François-Xavier Standaert, Nicolas Veyrat-Charvillon, Elisabeth Oswald, Benedikt Gierlichs, Marcel Medwed, Markus Kasper, Stefan Mangard: *The World Is Not Enough: Another Look on Second-Order DPA*. ASIACRYPT 2010: 112-129
[CFOS21] Gaëtan Cassiers, Sebastian Faust, Maximilian Orlt, François-Xavier Standaert: *Towards Tight Random Probing Security.* CRYPTO (3) 2021: 185-214
[VMKS12] Nicolas Veyrat-Charvillon, Marcel Medwed, Stéphanie Kerckhof, François-Xavier Standaert:
Shuffling against Side-Channel Attacks: A Comprehensive Study with Cautionary Note. ASIACRYPT 2012: 740-757
[CS19] Gaëtan Cassiers, François-Xavier Standaert: *Towards Globally Optimized Masking: From Low Randomness to Low Noise Rate or Probe Isolating Multiplications with Reduced Randomness and Security against Horizontal Attacks*. IACR Trans. Cryptogr. Hardw. Embed. Syst. 2019(2): 162-198 (2019)
[A+22] Melissa Azouaoui, Olivier Bronchain, Vincent Grosso, Kostas Papagiannopoulos: *Bitslice Masking and Improved Shuffling: How and When to Mix Them in Software?* IACR Trans. Cryptogr. Hardw. Embed. Syst. 2022(2): 140-165 (2022)
[A+20] Melissa Azouaoui, Davide Bellizia, Ileana Buhan, Nicolas Debande, Sébastien Duval, Christophe Giraud, Éliane Jaulmes, François Koeune, Elisabeth Oswald, François-Xavier Standaert, Carolyn Whitnall: A Systematic Appraisal of Side Channel Evaluation Strategies. SSR 2020: 46-66