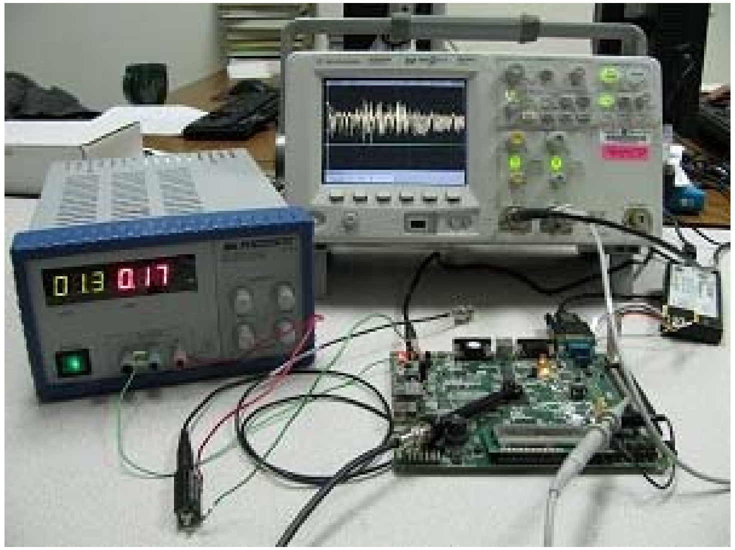
How to Certify the Leakage of a Chip?



F. Durvaux, ***F.-X. Standaert***, N. Veyrat-Charvillon

UCL Crypto Group, Belgium

EUROCRYPT 2014, Copenhagen, Denmark

Problem statement

Evaluation / certification of leaking devices

Motivation

- We currently lack formal approaches to “prove” the security of cryptographic implementations
  - Despite progresses in leakage-resilience

Motivation

- We currently lack formal approaches to "prove" the security of cryptographic implementations
  - Despite progresses in leakage-resilience

- The secure smart cards in your pockets usually go through the process of evaluation/certification
  - i.e. they are sent to a lab for evaluation and come back with a "security stamp" (A,B,C, …)

Motivation

- We currently lack formal approaches to "prove" the security of cryptographic implementations
  - Despite progresses in leakage-resilience

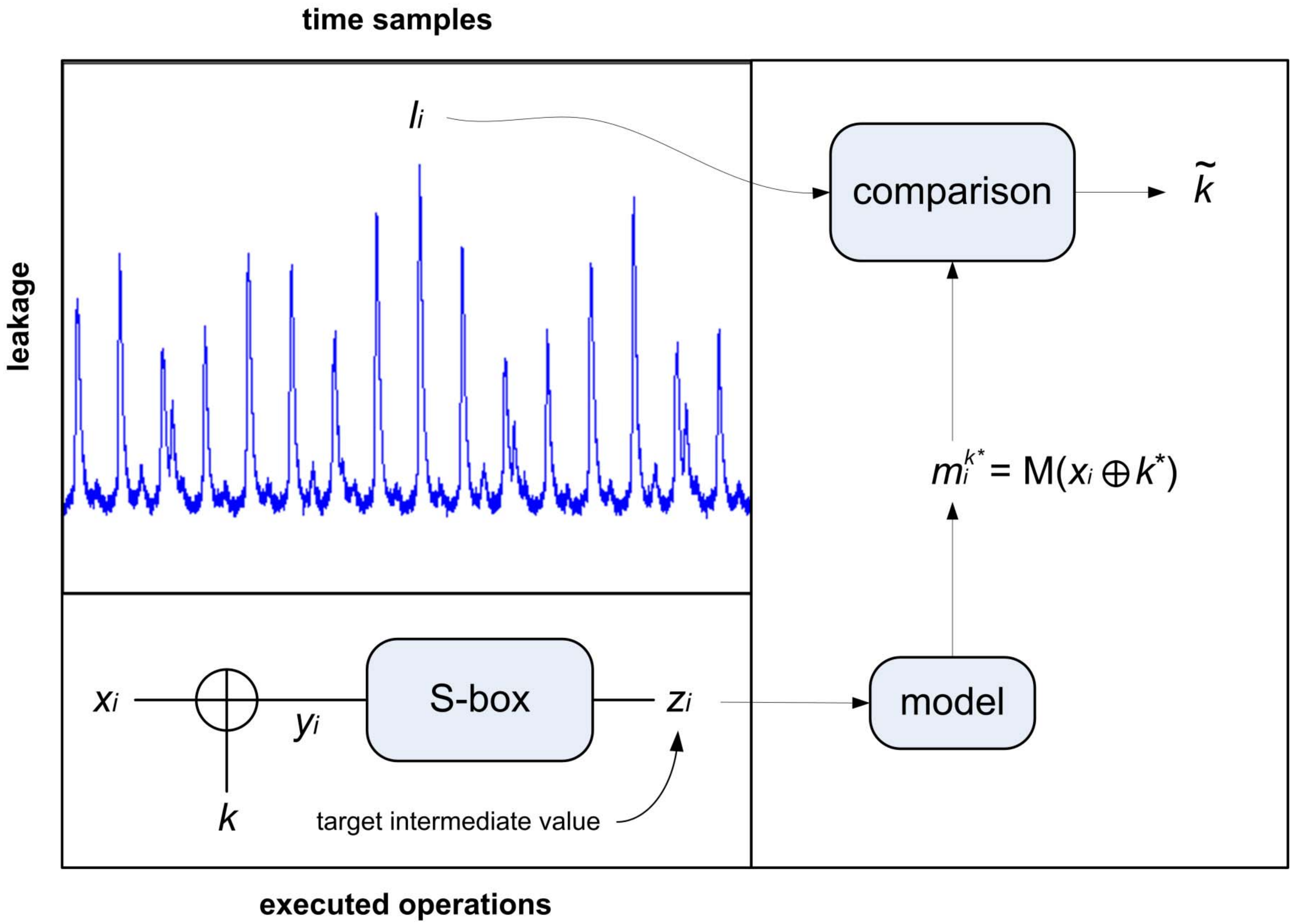
- The secure smart cards in your pockets usually go through the process of evaluation/certification
  - i.e. they are sent to a lab for evaluation and come back with a "security stamp" (A,B,C, …)

- This talk is about how to perform evaluations

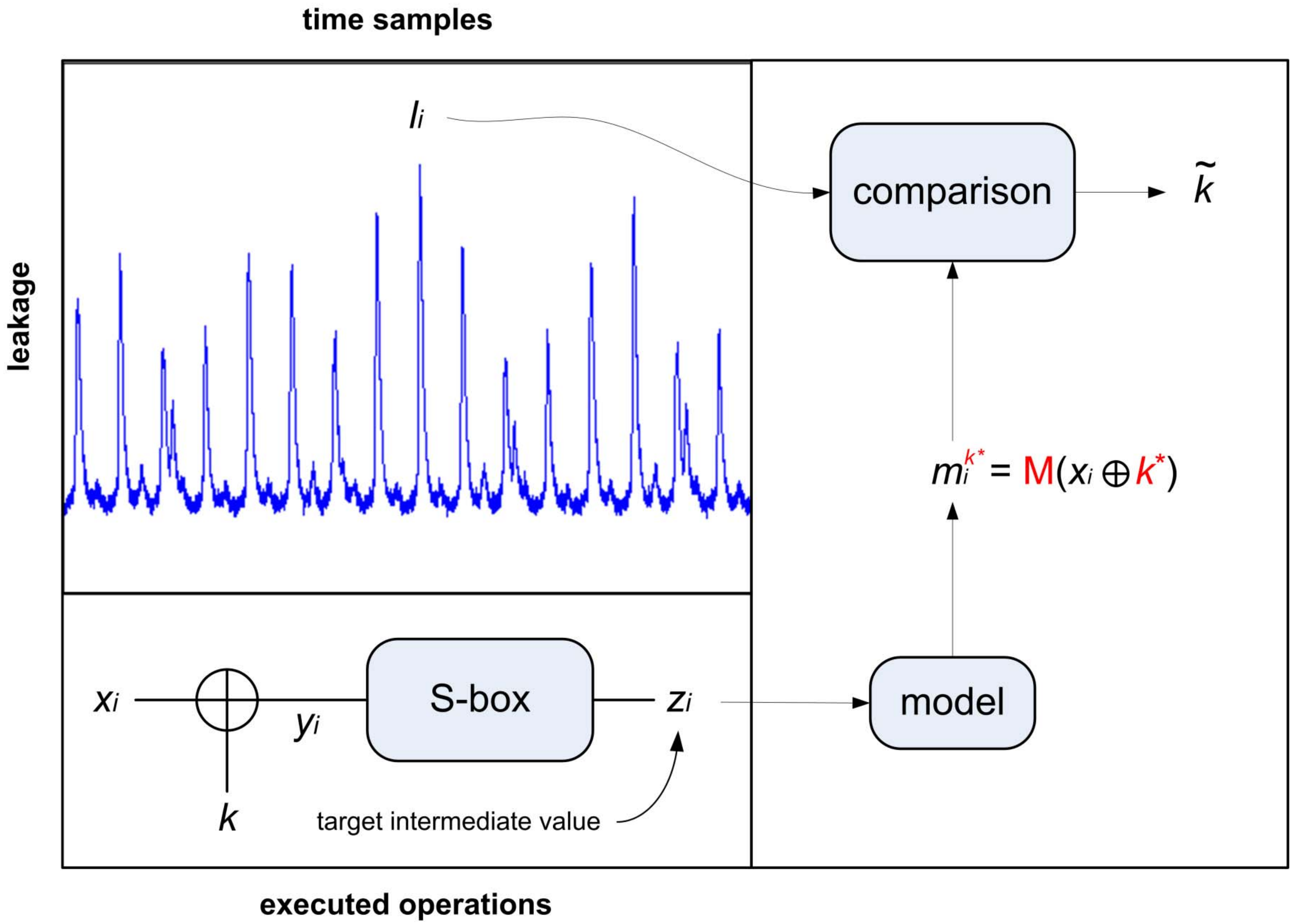
=> Quantified levels rather than hard to interpret letters

(≈ compute the ε's in proofs of leakage-resilience)

Side-channel attacks



Side-channel attacks



How to evaluate SCAs?

- Ideally, we should consider *worst-case* attacks

How to evaluate SCAs?

- Ideally, we should consider *worst-case* attacks

- But side-channel attacks rely on hypotheses
  - on the target piece of key (*useful*)
  - and on the leakage model (*useless*)

How to evaluate SCAs?

- Ideally, we should consider *worst-case* attacks

- But side-channel attacks rely on hypotheses
  - on the target piece of key (*useful*)
  - and on the leakage model (*useless*)

=> Worst-case evaluations require a *perfect* model

How to evaluate SCAs?

- Ideally, we should consider *worst-case* attacks

- But side-channel attacks rely on hypotheses
  - on the target piece of key (*useful*)
  - and on the leakage model (*useless*)

=> Worst-case evaluations require a *perfect* model

- Problem: such a (*physical*) model is unknown!

How to evaluate SCAs?

- Ideally, we should consider *worst-case* attacks

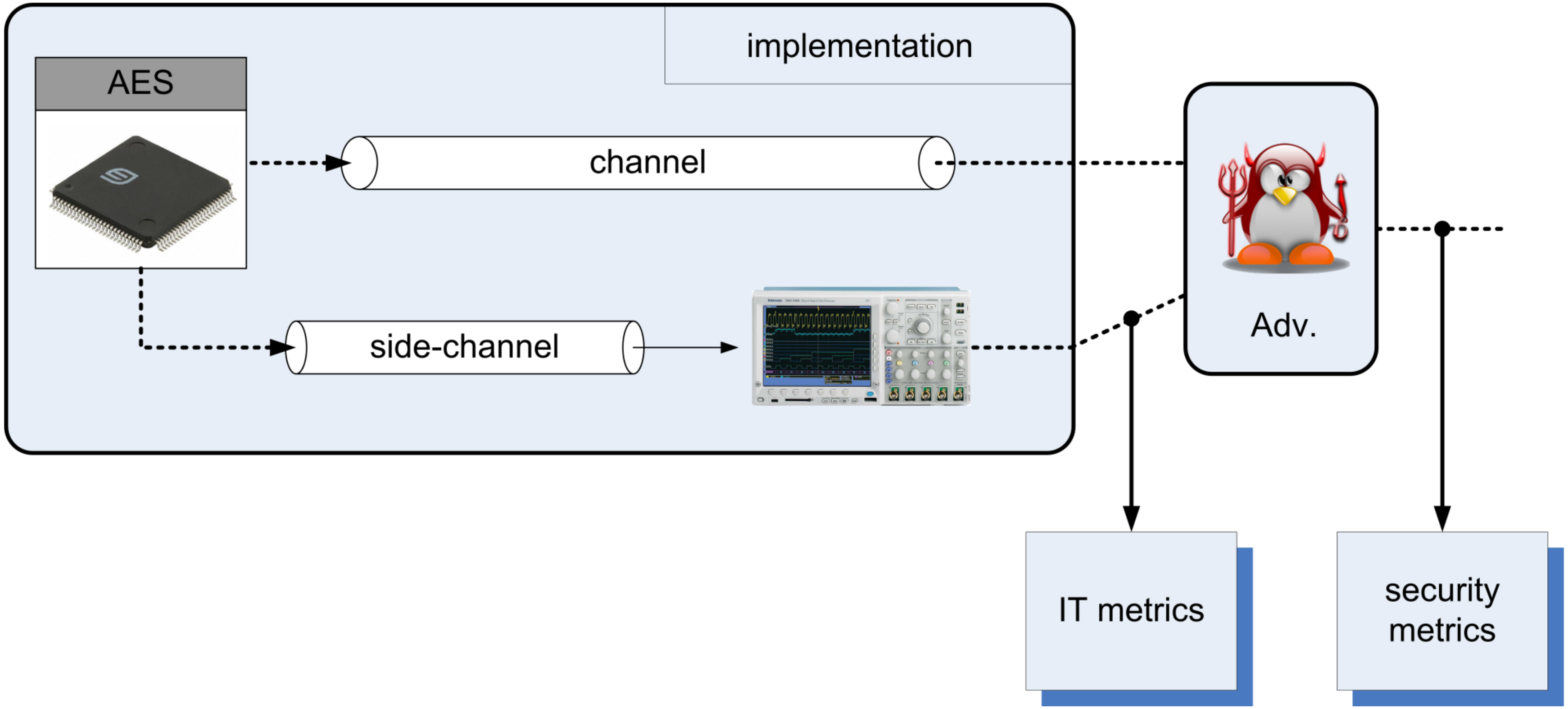
- But side-channel attacks rely on hypotheses
  - on the target piece of key (*useful*)
  - and on the leakage model (*useless*)

=> Worst-case evaluations require a *perfect* model

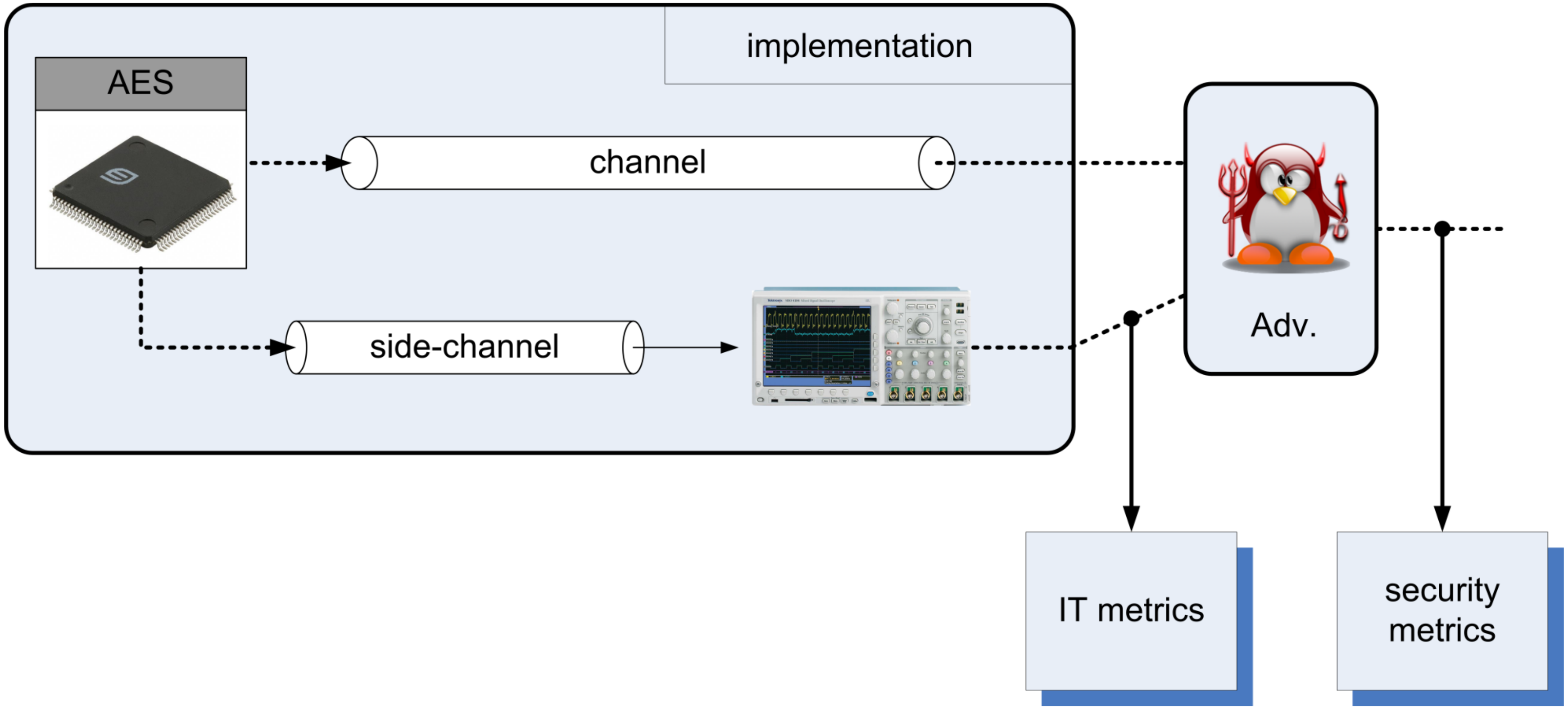
- Problem: such a (*physical*) model is unknown!

- This talk: *how good is my leakage model*?

Background: EC09 framework [1]

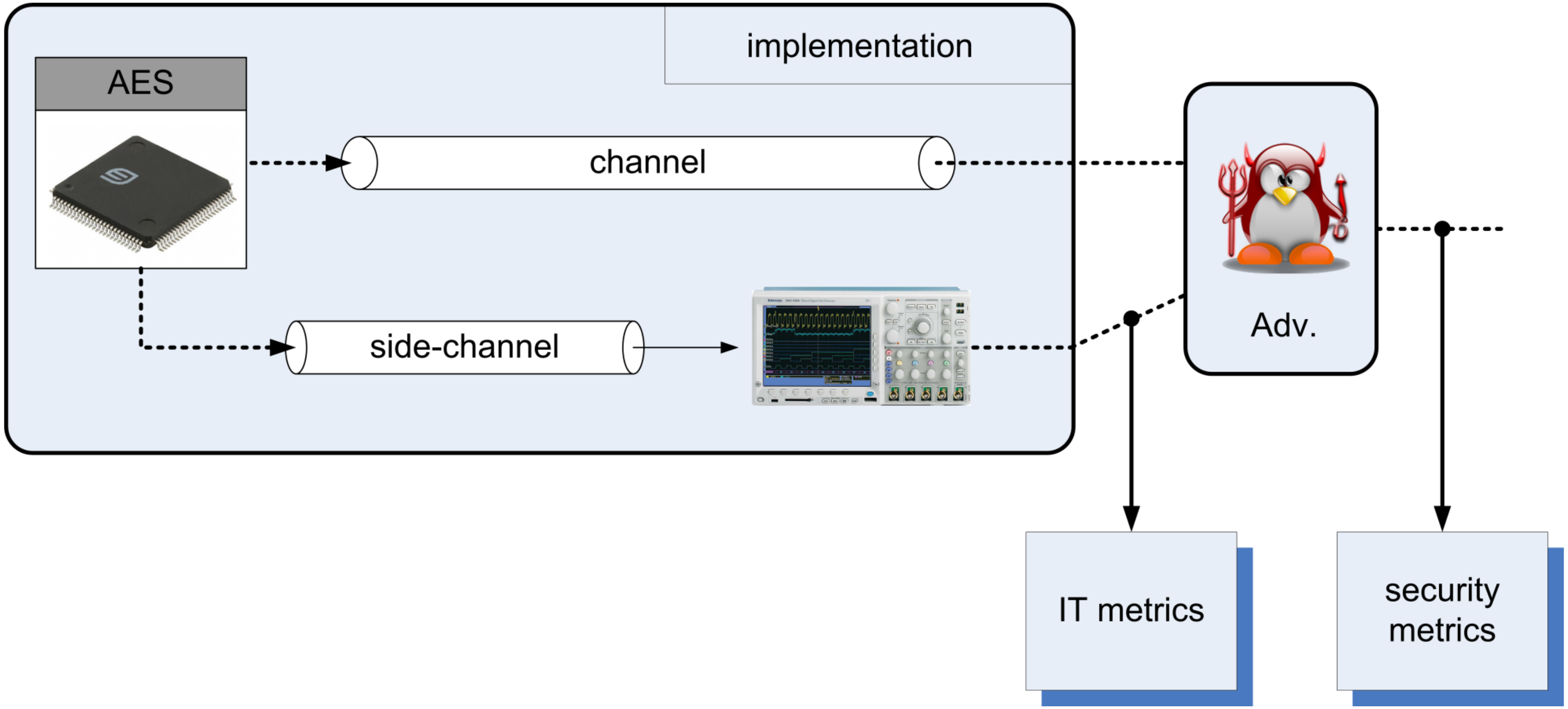


Background: EC09 framework [1]



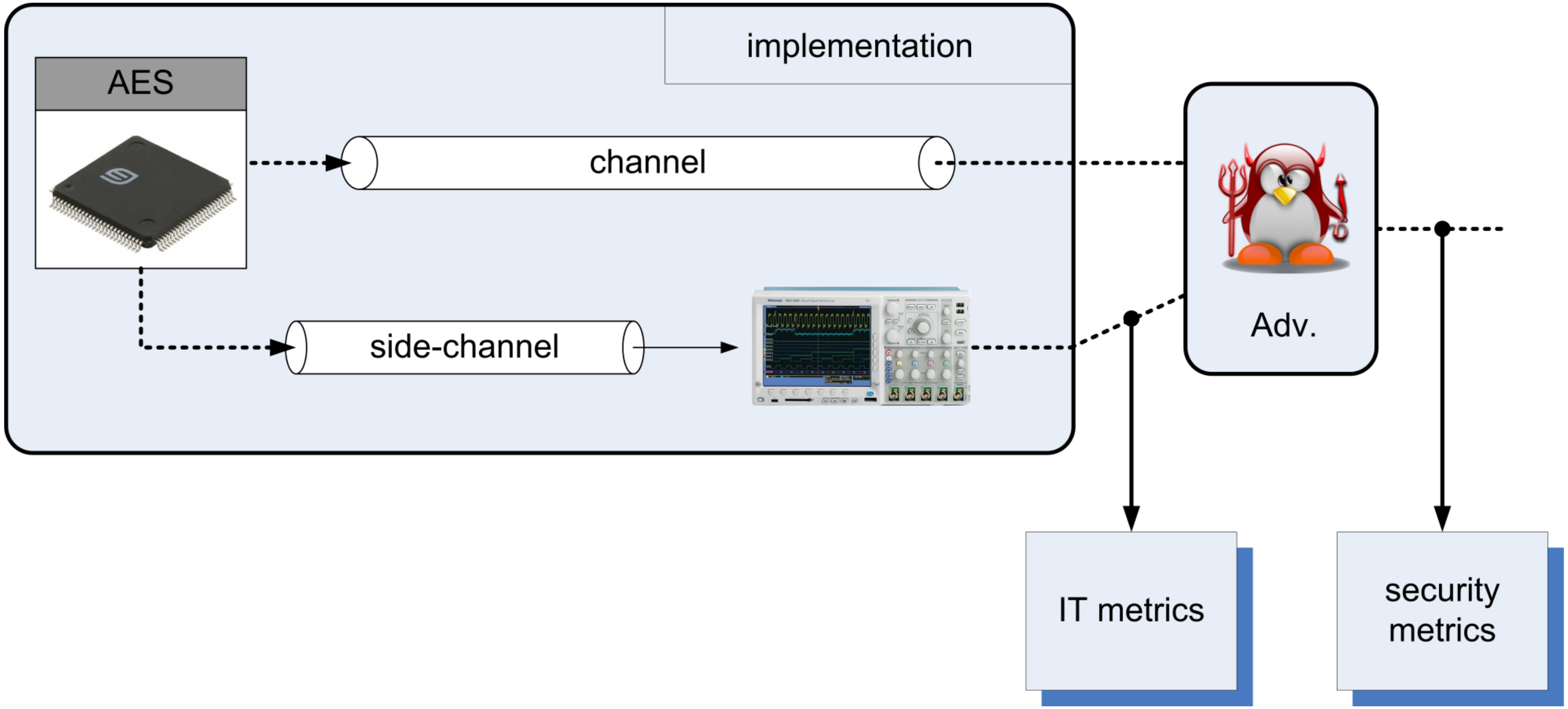
- Problem: estimating (e.g.) the mutual information between arbitrary distributions is notoriously hard!

Background: EC09 framework [1]



- Good news: side-channel attacks need a model
  - i.e. an estimation of the leakage distribution

Background: EC09 framework [1]



- Main idea: estimate the mutual information from the “best available” model (*practical worst case*)

Definition

- Information leakage on the secret key

$$\mathrm{H}[K] - \sum_{k} \Pr[k] \sum_{l} \Pr_{chip}[l|k] \,.\log_2 \widehat{\Pr}_{model}[k|l]$$

- where $\widehat{\Pr}_{model}[k|l]$ is obtained by profiling
- and $\Pr_{chip}[l|k]$ is unknown but can be sampled

Two cases can happen [2]

- Case #1 (ideal): perfect profiling phase
- i.e. $\widehat{\Pr}_{model}\,[l|k] = \Pr_{chip}\,[l|k]$

$$\widehat{\mathrm{MI}}(K;L) = \mathrm{H}[K] - \sum_{k} \Pr[k] \sum_{l} \Pr_{chip}\,[l|k]\,.\log_2 \Pr_{chip}\,[k|l]$$

Two cases can happen [2]

- Case #1 (ideal): perfect profiling phase
- i.e. $\widehat{\Pr}_{model}\,[l|k] = \Pr_{chip}\,[l|k]$

$$\widehat{\mathrm{MI}}(K;L) = \mathrm{H}[K] - \sum_{k} \Pr[k] \sum_{l} \Pr_{chip}\,[l|k]\,.\log_2 \Pr_{chip}\,[k|l]$$

- Case #2 (actual): bounded profiling phase
- i.e. $\widehat{\Pr}_{model}\,[l|k] \neq \Pr_{chip}\,[l|k]$

$$\widehat{\mathrm{PI}}(K;L) = \mathrm{H}[K] - \sum_{k} \Pr[k] \sum_{l} \Pr_{chip}\,[l|k]\,.\log_2 \widehat{\Pr}_{model}\,[k|l]$$

Evaluation problem statement

- What is the distance between the MI and the PI?
  (i.e. *how good is my leakage model*?)

Evaluation problem statement

- What is the distance between the MI and the PI?
  (i.e. *how good is my leakage model*?)

- Difficult since the leakage function is unknown

=> Impossible to compute this distance directly!

Evaluation problem statement

- What is the distance between the MI and the PI?
  (i.e. *how good is my leakage model*?)

- Difficult since the leakage function is unknown

=> Impossible to compute this distance directly!

- Our result: we show that indirect approaches allow answering the question quite rigorously

Evaluation problem statement

- What is the distance between the MI and the PI?
  (i.e. *how good is my leakage model*?)

- Difficult since the leakage function is unknown

=> Impossible to compute this distance directly!

- Our result: we show that indirect approaches allow answering the question quite rigorously

- Main idea: quantify the different model errors!

First question: estimation errors

Has my model converged?

Cross-validation

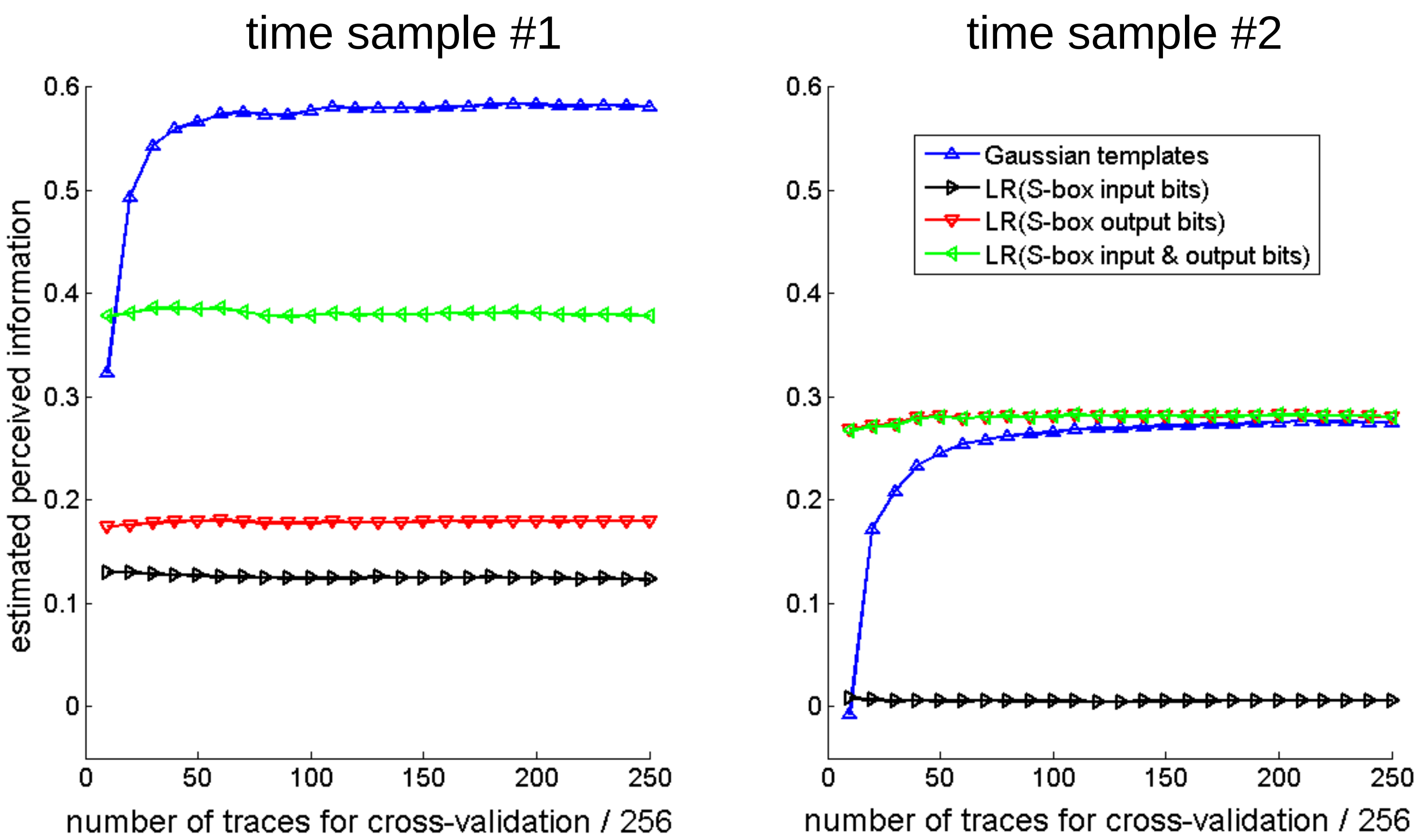
- Split traces in 10 (non-overlapping) sets, use 9/10th for profiling, 1/10th for estimating the PI
- Repeat 10 times to get average & spread

Cross-validation

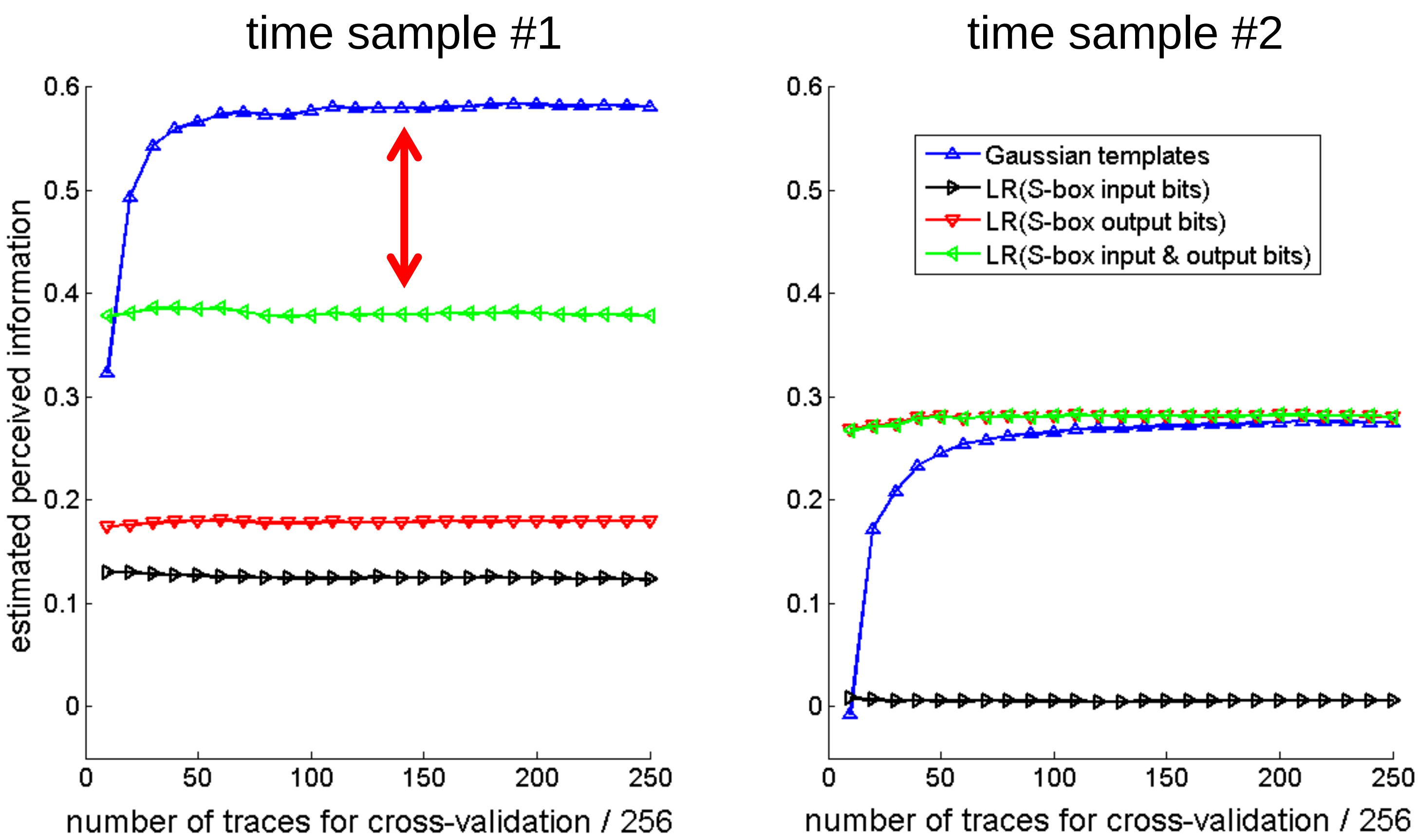
- Split traces in 10 (non-overlapping) sets, use 9/10th for profiling, 1/10th for estimating the PI
- Repeat 10 times to get average & spread

- Example of models
  - *Gaussian templates*: estimate one Gaussian distribution per value of $x_i \oplus k_i$
  - *Linear regression*: approximate L(x_i,k_i) with a linear combination of basis elements
    - e.g. the S-box input & output bits

Cross-validation results

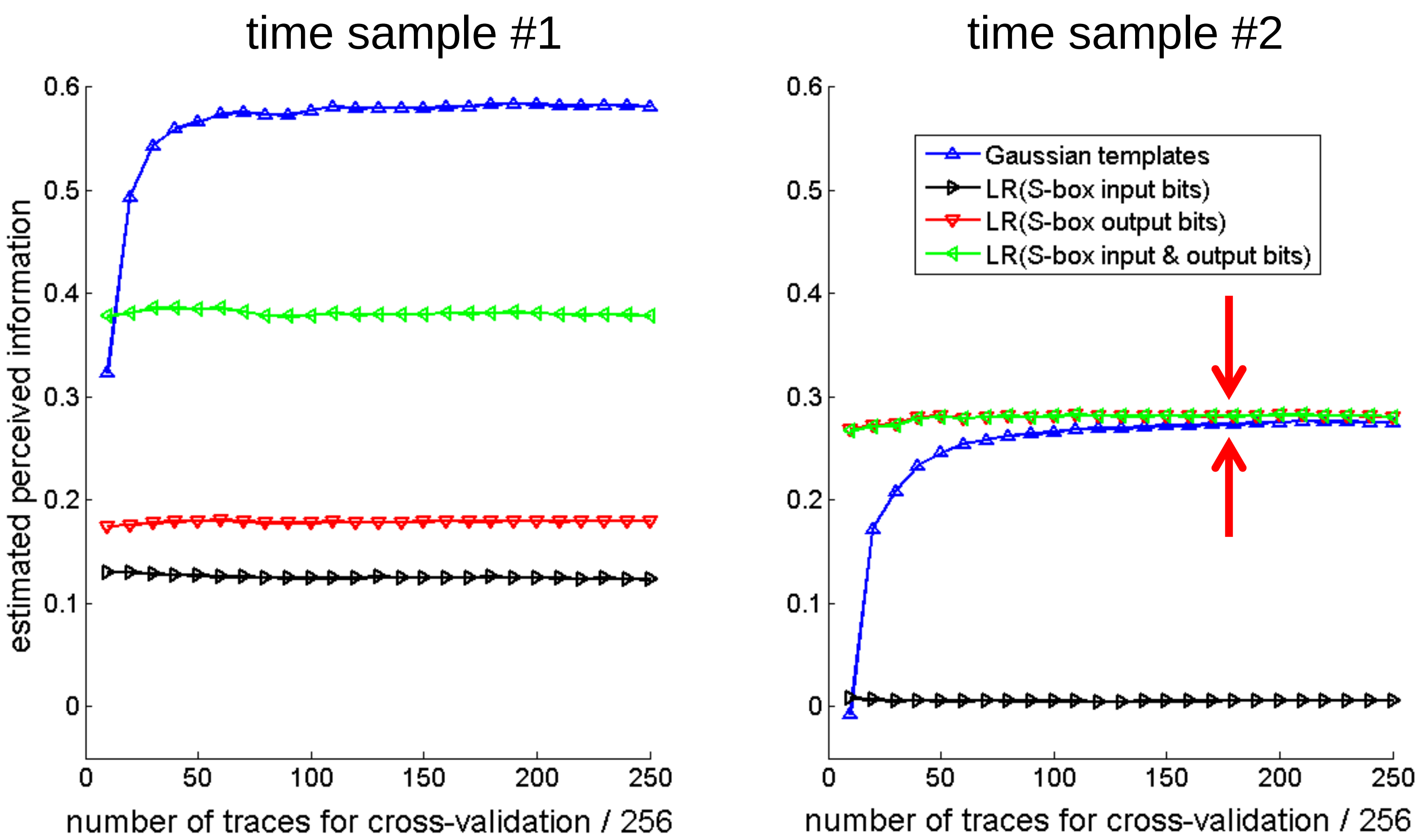


Cross-validation results



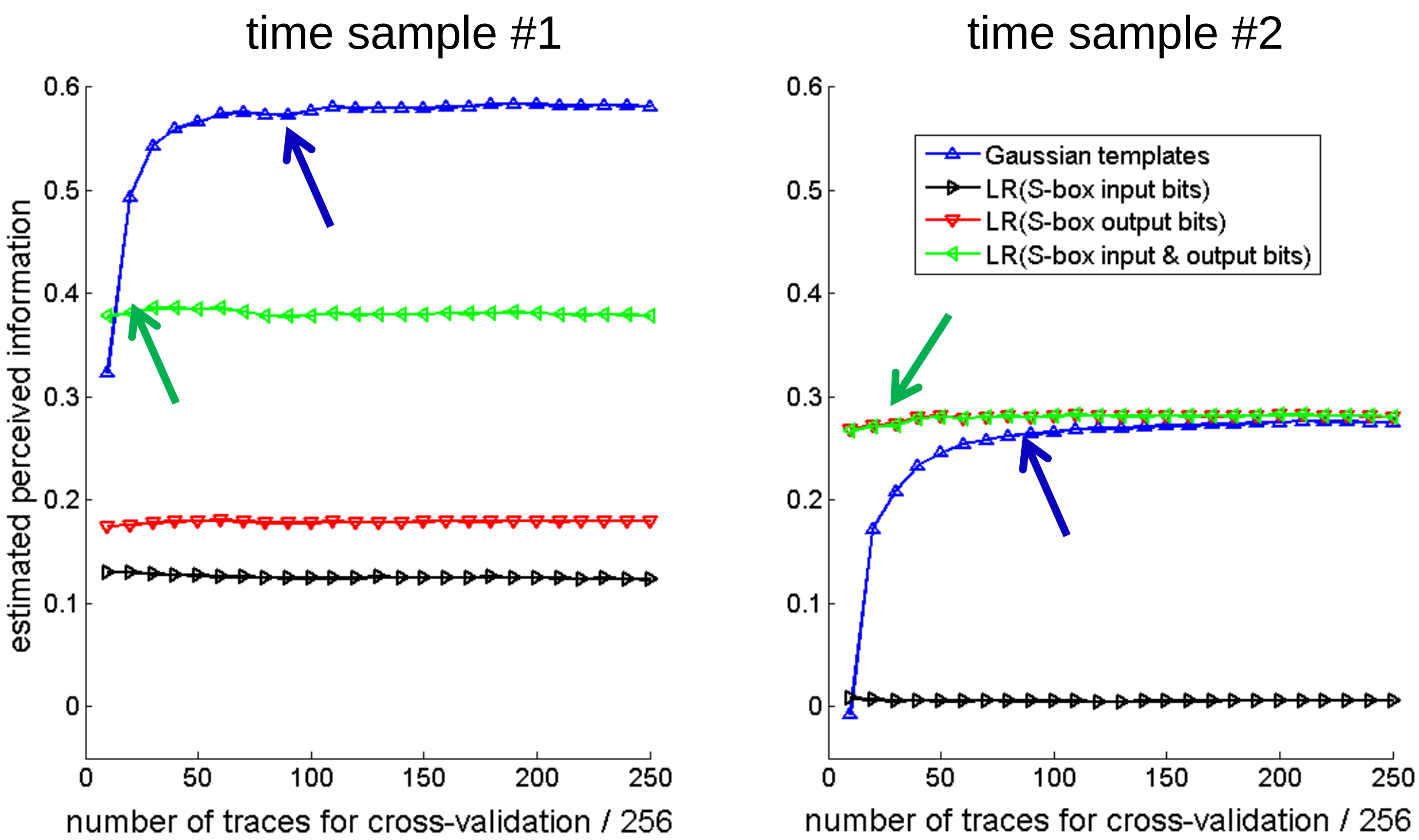
- Gaussian templates more informative for t1

Cross-validation results



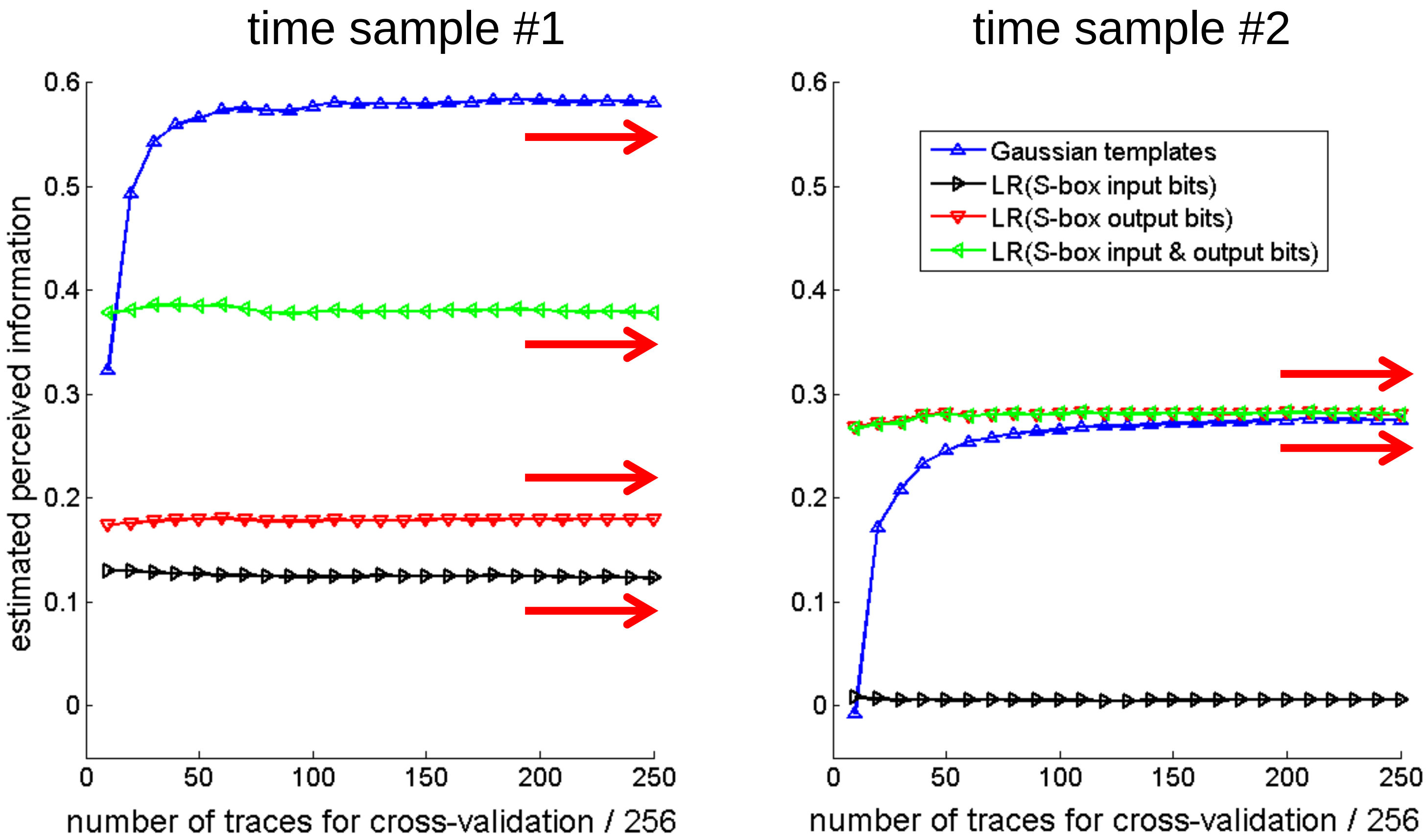
- Linear basis with S-box output bits sufficient for t2

Cross-validation results



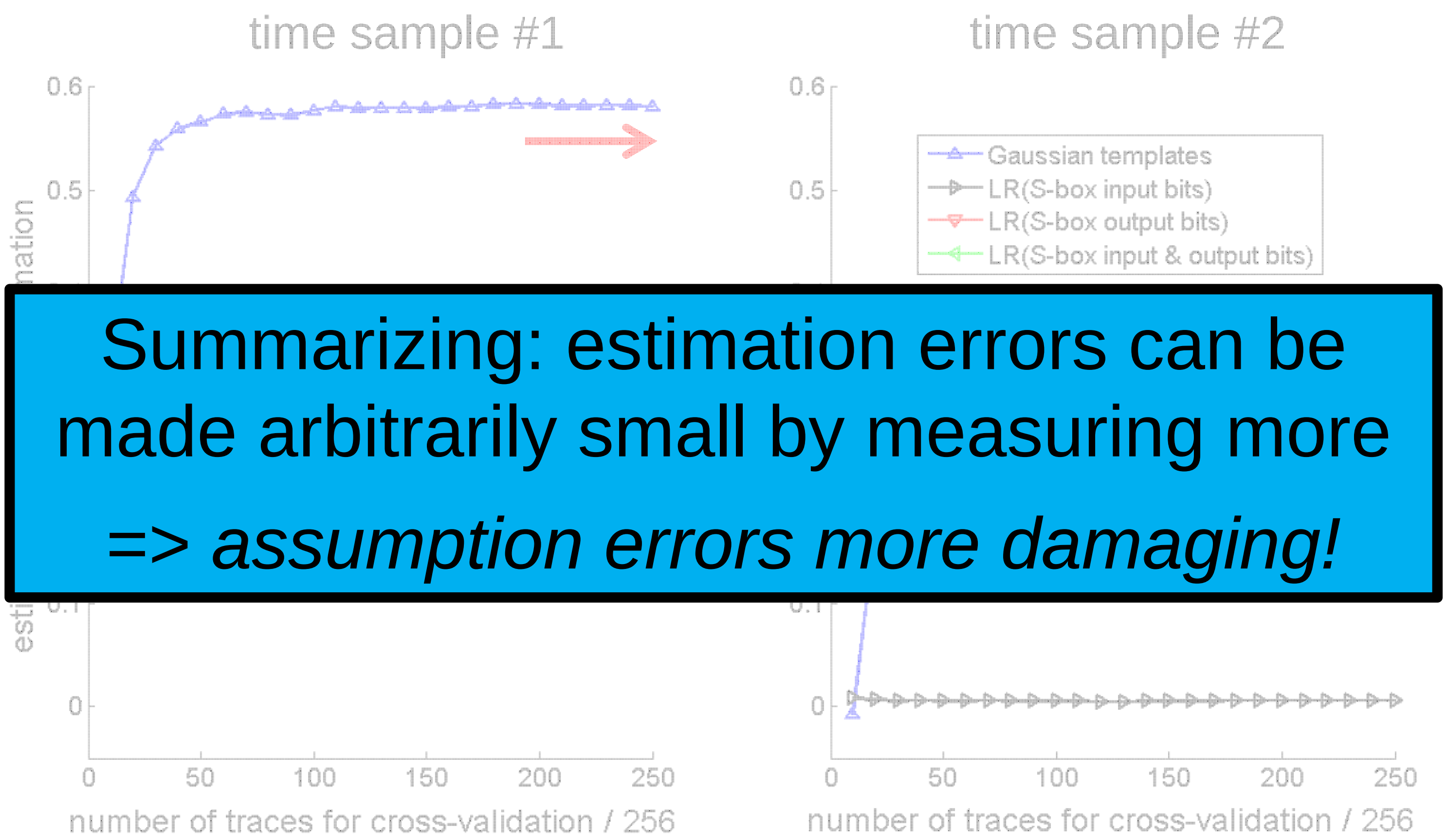
- Estimation of Gaussian templates more expensive

Cross-validation results



- All models have converged after ~50,000 traces

Cross-validation results



Summarizing: estimation errors can be made arbitrarily small by measuring more

=> *assumption errors more damaging!*

- All models have converged after ~50,000 traces

Second question: assumption errors

Is my model good enough?

(PART I: conditioned on the # of measurements)

Step a. Detecting assumption errors

- Fact: two multidimensional distributions $\mathcal{F}$ and $\mathcal{G}$ are equal if the variables X~$\mathcal{F}$ and Y~$\mathcal{G}$ generate identical distributions for the distance D(.,.)

Step a. Detecting assumption errors

- Fact: two multidimensional distributions $\mathcal{F}$ and $\mathcal{G}$ are equal if the variables X~$\mathcal{F}$ and Y~$\mathcal{G}$ generate identical distributions for the distance D(.,.)

- We can compute the simulated distance

$$f_{sim}(d) = \Pr[L_1 - L_2 \leq d \mid L_1, L_2 \sim \widehat{\Pr}_{model}]$$

Step a. Detecting assumption errors

- Fact: two multidimensional distributions $\mathcal{F}$ and $\mathcal{G}$ are equal if the variables X~$\mathcal{F}$ and Y~$\mathcal{G}$ generate identical distributions for the distance D(.,.)

- We can compute the simulated distance

$$f_{sim}(d) = \Pr[L_1 - L_2 \leq d \mid L_1, L_2 \sim \widehat{\Pr}_{model}]$$

- And the sampled distance

$$\hat{g}_N(d) = \Pr[l_1 - l_2 \leq d \mid l_1 \overset{N}{\Leftarrow} \widehat{\Pr}_{model}, l_2 \overset{N}{\Leftarrow} \Pr_{chip}]$$

Step a. Detecting assumption errors

- Fact: two multidimensional distributions $\mathcal{F}$ and $\mathcal{G}$ are equal if the variables X~$\mathcal{F}$ and Y~$\mathcal{G}$ generate identical distributions for the distance D(.,.)

- We can compute the simulated distance

$$f_{sim}(d) = \Pr[L_1 - L_2 \le d \mid L_1, L_2 \sim \widehat{\Pr}_{model}]$$

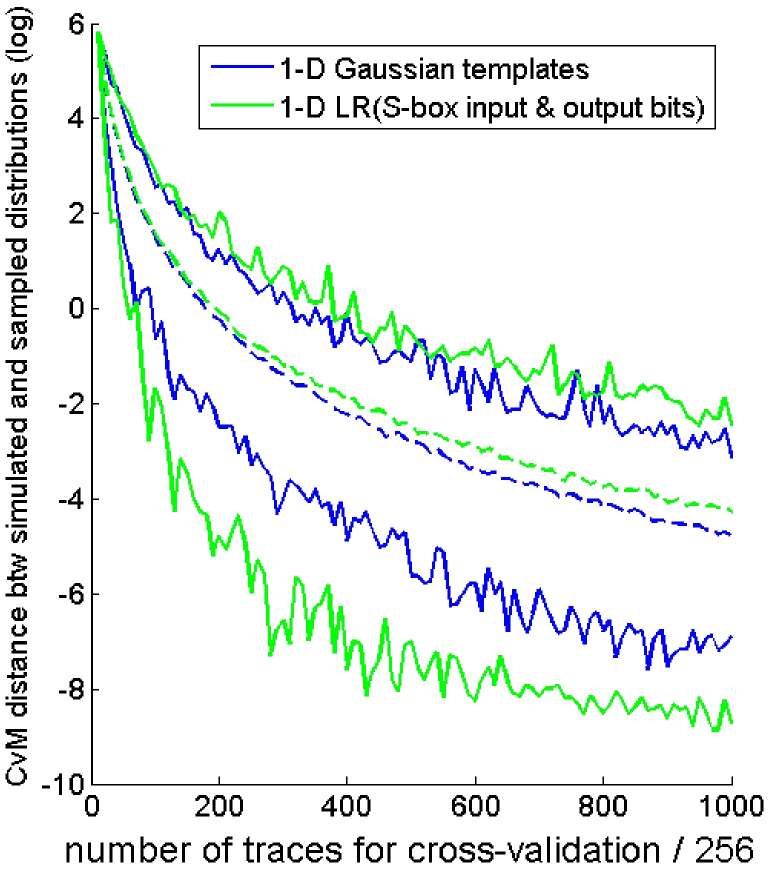
- And the sampled distance

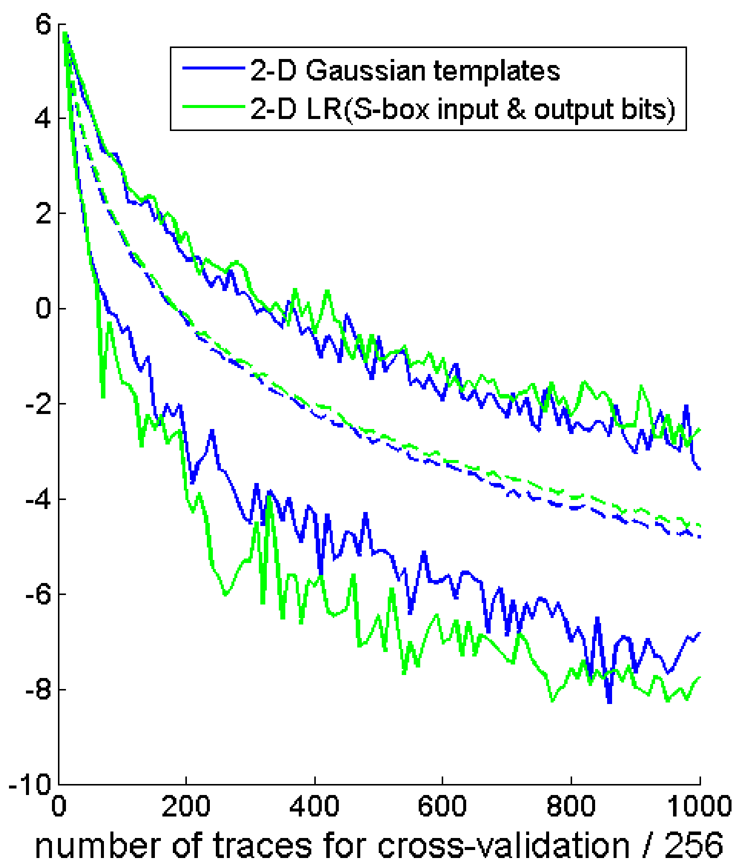
$$\hat{g}_N(d) = \Pr[l_1 - l_2 \le d \mid l_1 \overset{N}{\Leftarrow} \widehat{\Pr}_{model}, l_2 \overset{N}{\Leftarrow} \Pr_{chip}]$$

- And test their CvM divergence

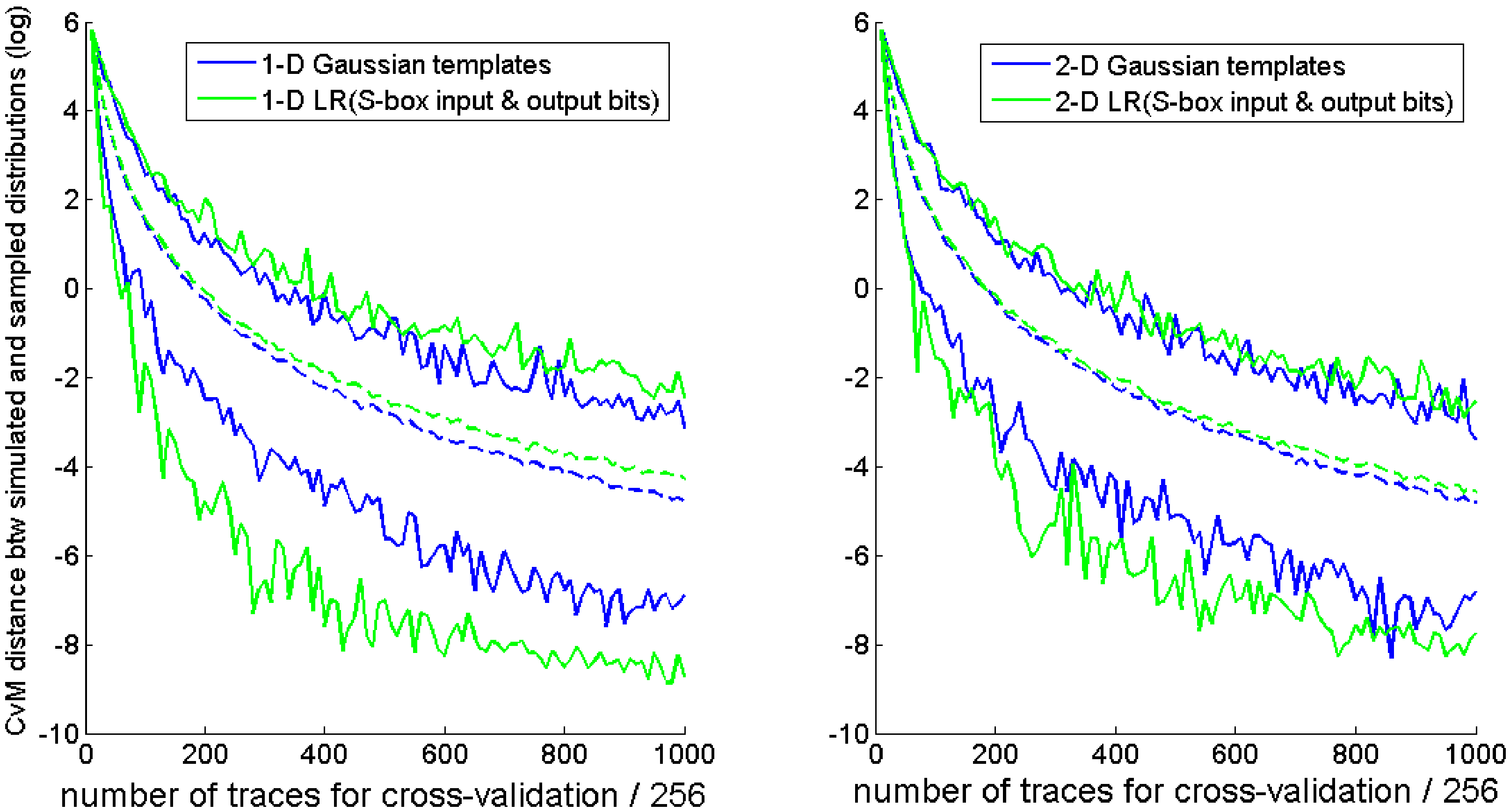
$$\widehat{\mathrm{CvM}}\,(f_{sim}, \hat{g}_N) = \int [f_{sim}(x) - \hat{g}_N(x)]^2 dx$$

With cross-validation again, we obtain



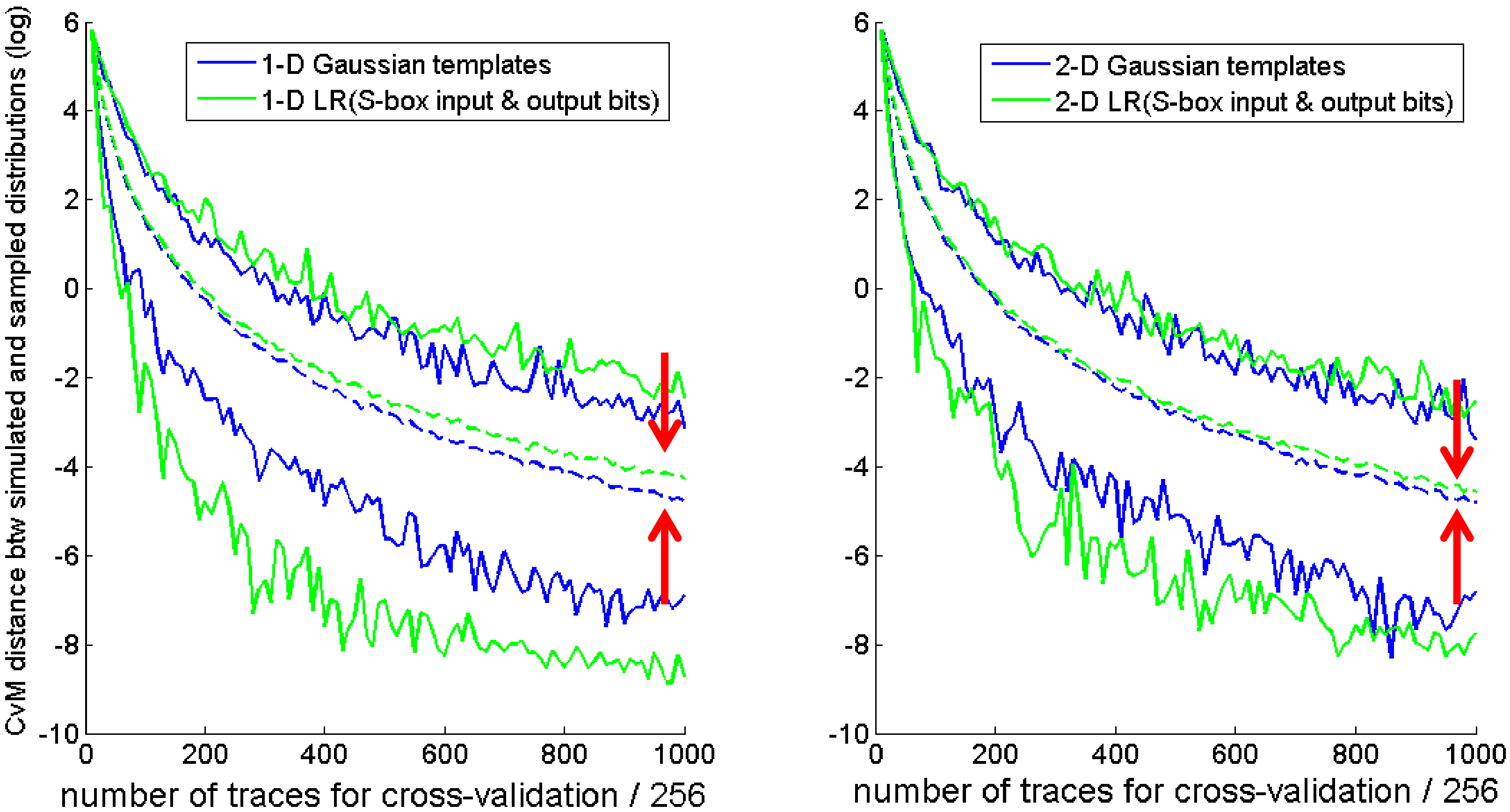


With cross-validation again, we obtain



- Any incorrect assumption => CvM saturates

With cross-validation again, we obtain



- Are these models already saturating?

Step b. Estimation vs. assumption errors

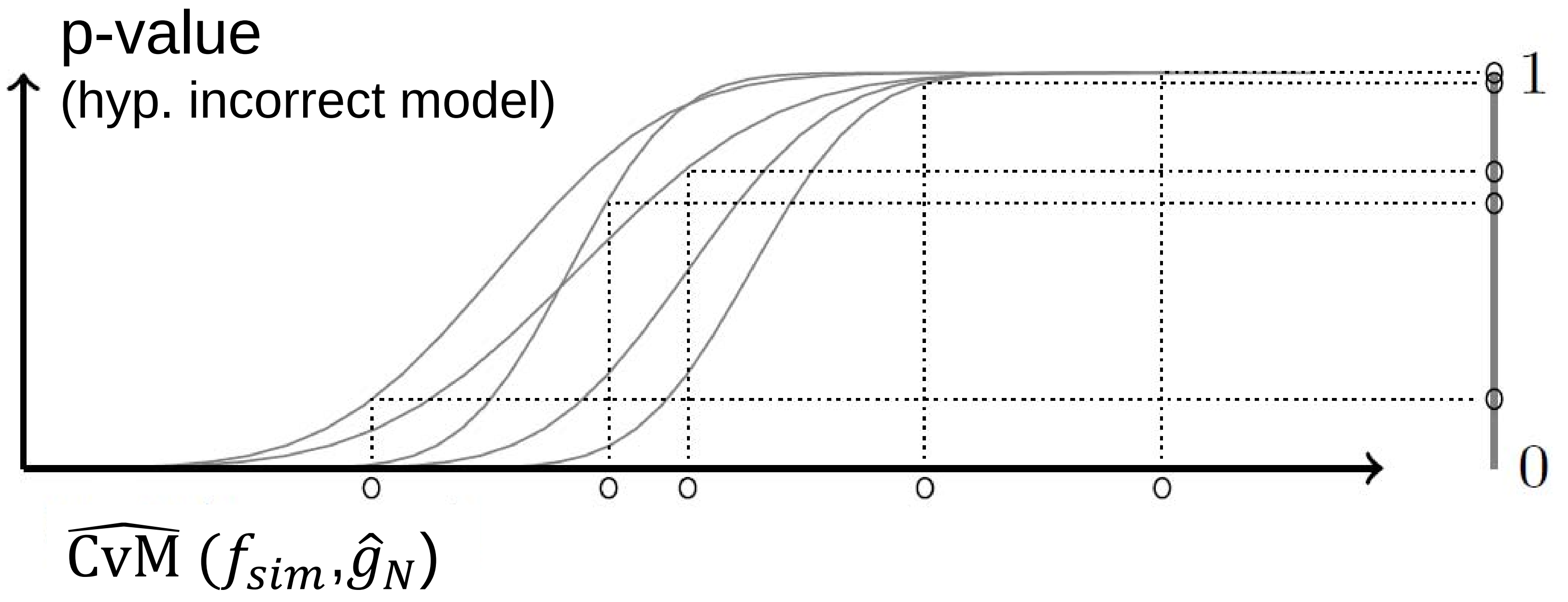
- Goal: try to detect when assumption errors become significant in front of estimation ones

Step b. Estimation vs. assumption errors

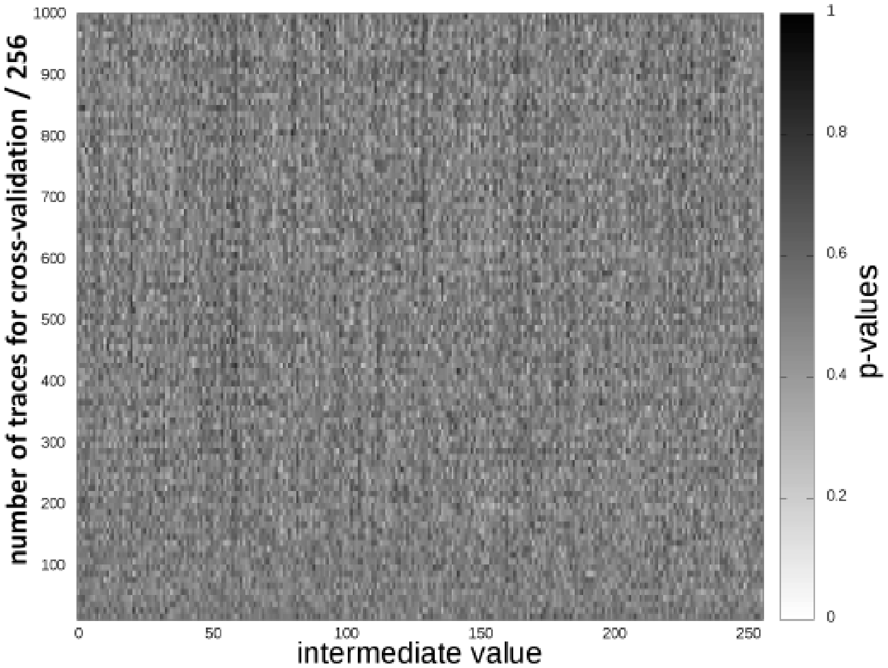
- Goal: try to detect when assumption errors become significant in front of estimation ones
  - Characterize the probability that a given model error can be explained by estimation issues

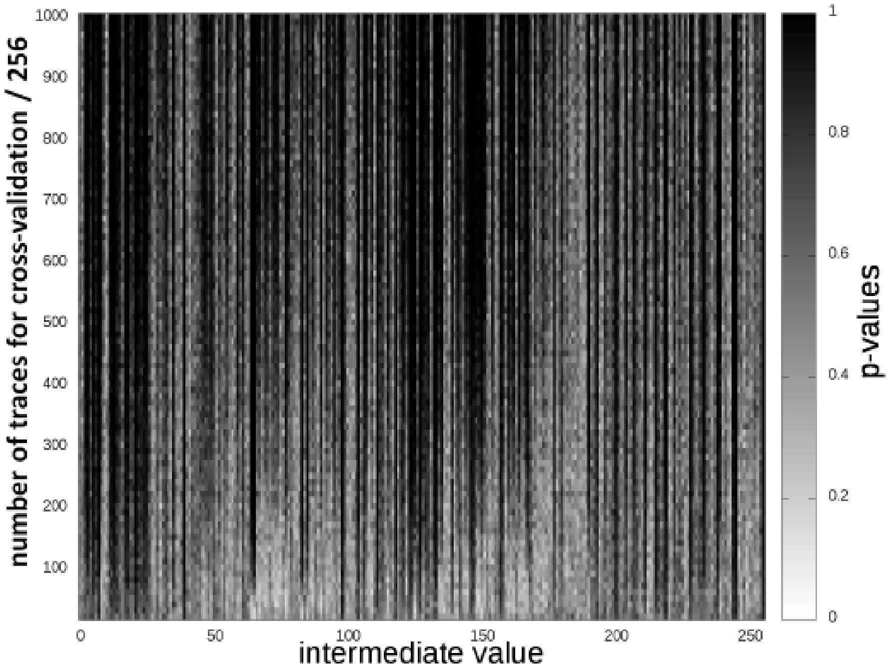
Step b. Estimation vs. assumption errors

- Goal: try to detect when assumption errors become significant in front of estimation ones
    - Characterize the probability that a given model error can be explained by estimation issues



Comparison of models

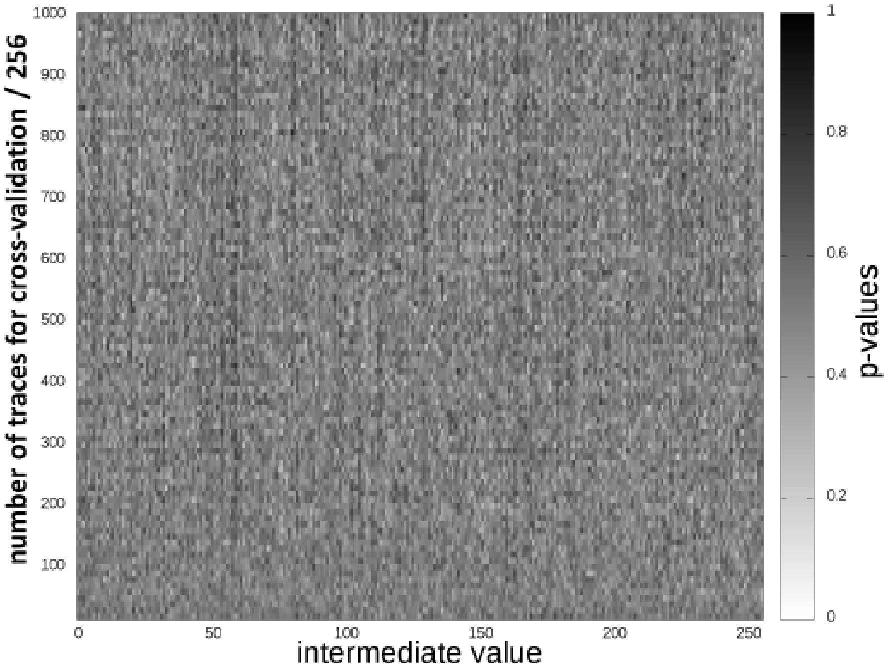




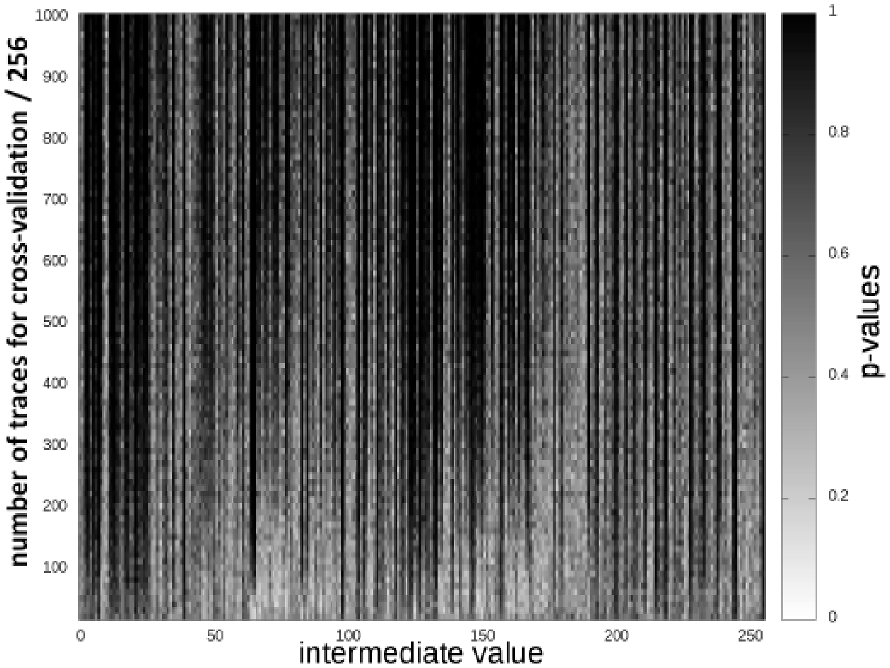
Gaussian templates

Linear regression

Comparison of models



Gaussian templates



Linear regression

=> Gaussian templates are good enough with up to 256,000 traces in the cross-validation set

Second question: assumption errors

Is my model good enough?

(PART II: independent of the # of measurements)

Towards leakage bounds?

- Say we do measure up to the point where we detect assumption errors for all our models
- Can we bound the MI – PI difference?

Towards leakage bounds?

- Say we do measure up to the point where we detect assumption errors for all our models
- Can we bound the MI – PI difference?

- Attempt: for N_{th} such that the assumption errors are not significant in front of estimation errors, try to “bound” the information loss by quantifying the (easier to compute) estimation error

Towards leakage bounds?

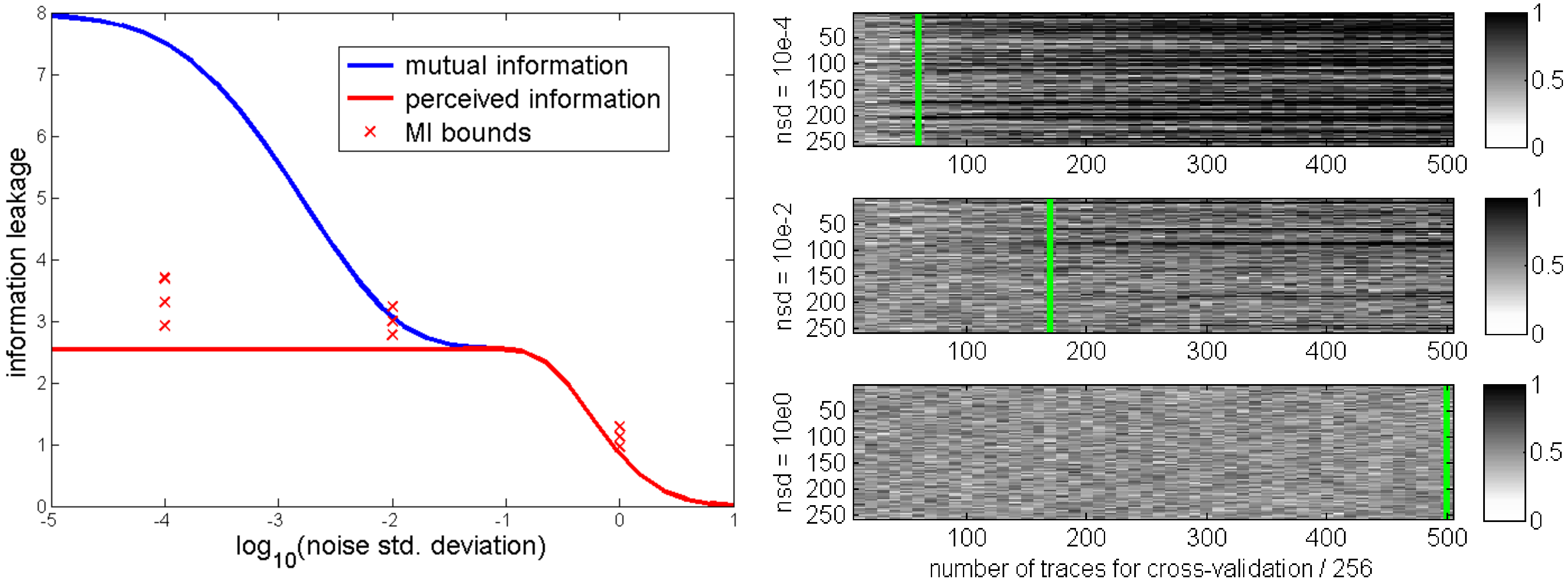
- Say we do measure up to the point where we detect assumption errors for all our models
- Can we bound the MI – PI difference?

- Attempt: for N_{th} such that the assumption errors are not significant in front of estimation errors, try to "bound" the information loss by quantifying the (easier to compute) estimation error

  - Hope: *assumption errors that are detected for smaller N_{th}'s should be larger in some sense*

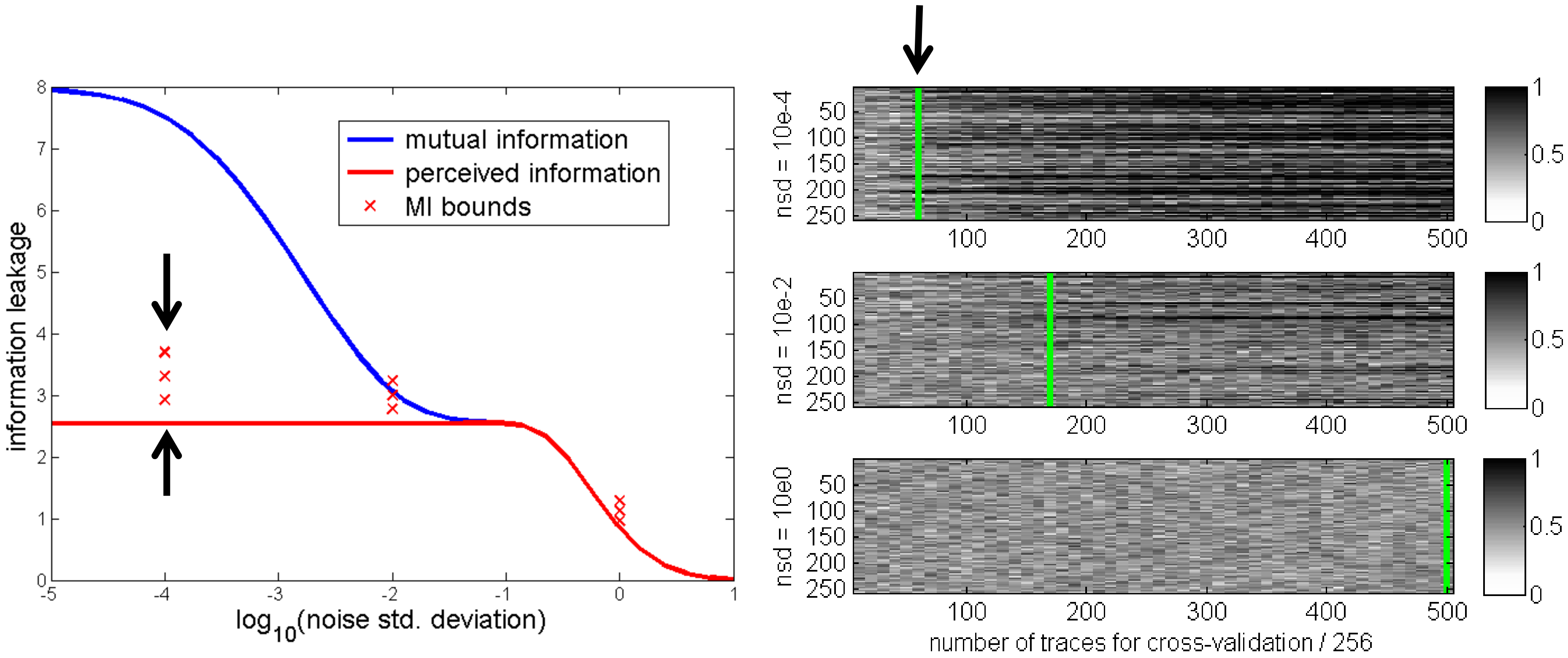
So far: counterexamples

- Mathematically generated leakages analyzed with LR (9-element basis) for different noise levels



So far: counterexamples

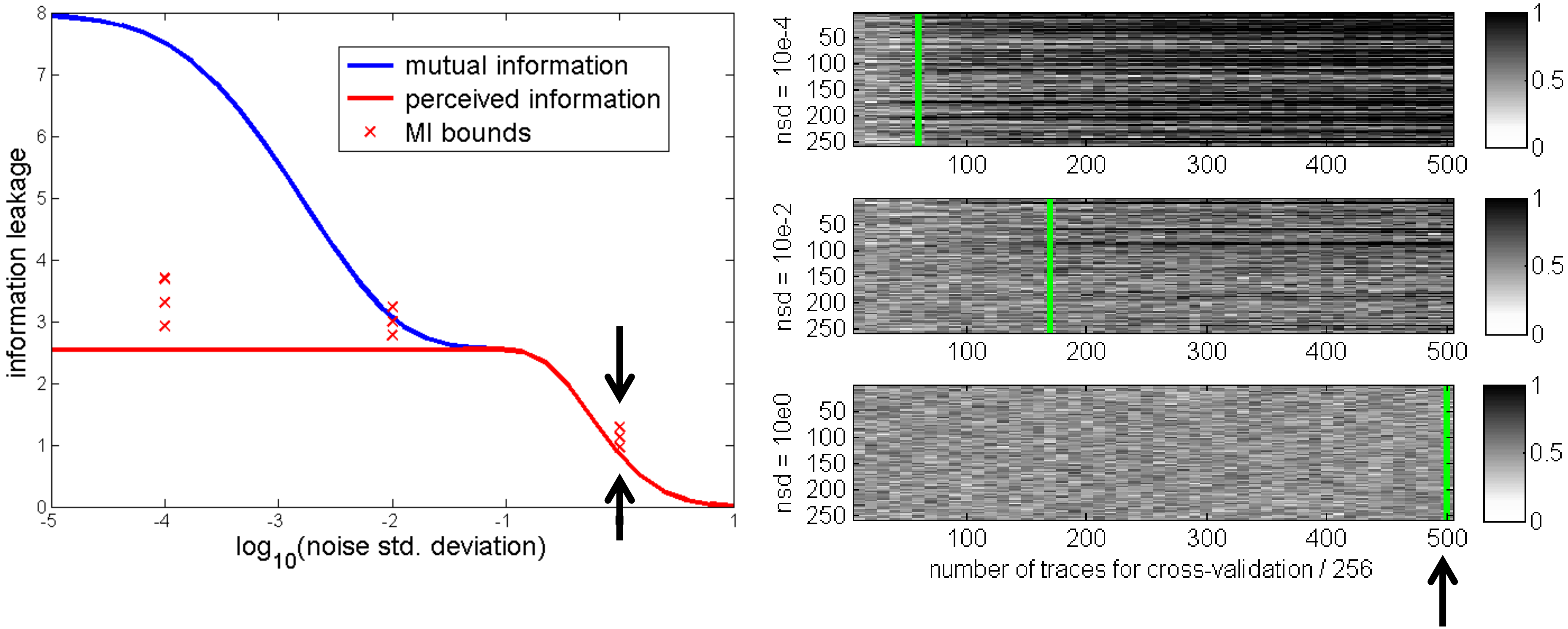
- Mathematically generated leakages analyzed with LR (9-element basis) for different noise levels



- Bound too optimistic for low noise levels

So far: counterexamples

- Mathematically generated leakages analyzed with LR (9-element basis) for different noise levels



- Bound too *pessimistic* for *large* noise levels

Main issues with such bounds

- The threshold for which assumption errors are detected (e.g. average p-value) is hard to set independent of the leakage distributions

Main issues with such bounds

- The threshold for which assumption errors are detected (e.g. average p-value) is hard to set independent of the leakage distributions

- Information bounds anyway become pessimistic as the noise increases (since the noise then dominates the assumption errors in the MSE)

Main issues with such bounds

- The threshold for which assumption errors are detected (e.g. average p-value) is hard to set independent of the leakage distributions
- Information bounds anyway become pessimistic as the noise increases (since the noise then dominates the assumption errors in the MSE)

There could be more positive results for certain distributions (*scope for further research*), meanwhile…

Pragmatic evaluation guidelines

- ***For a fixed number of measurements***
  (which is the case of all real-world evaluations)

Pragmatic evaluation guidelines

- ***For a fixed number of measurements***
  (which is the case of all real-world evaluations)

  - *If assumption errors are detected*: the loss of information due to an imprecise model is significant (i.e. the model can be improved)

Pragmatic evaluation guidelines

- ***For a fixed number of measurements***
  (which is the case of all real-world evaluations)

  - *If assumption errors are detected*: the loss of information due to an imprecise model is significant (i.e. the model can be improved)

  - *If assumption errors are not detected*: improving the model would not lead to better information extraction (since this improvement could not be distinguished due to the estimation errors)

Pragmatic evaluation guidelines

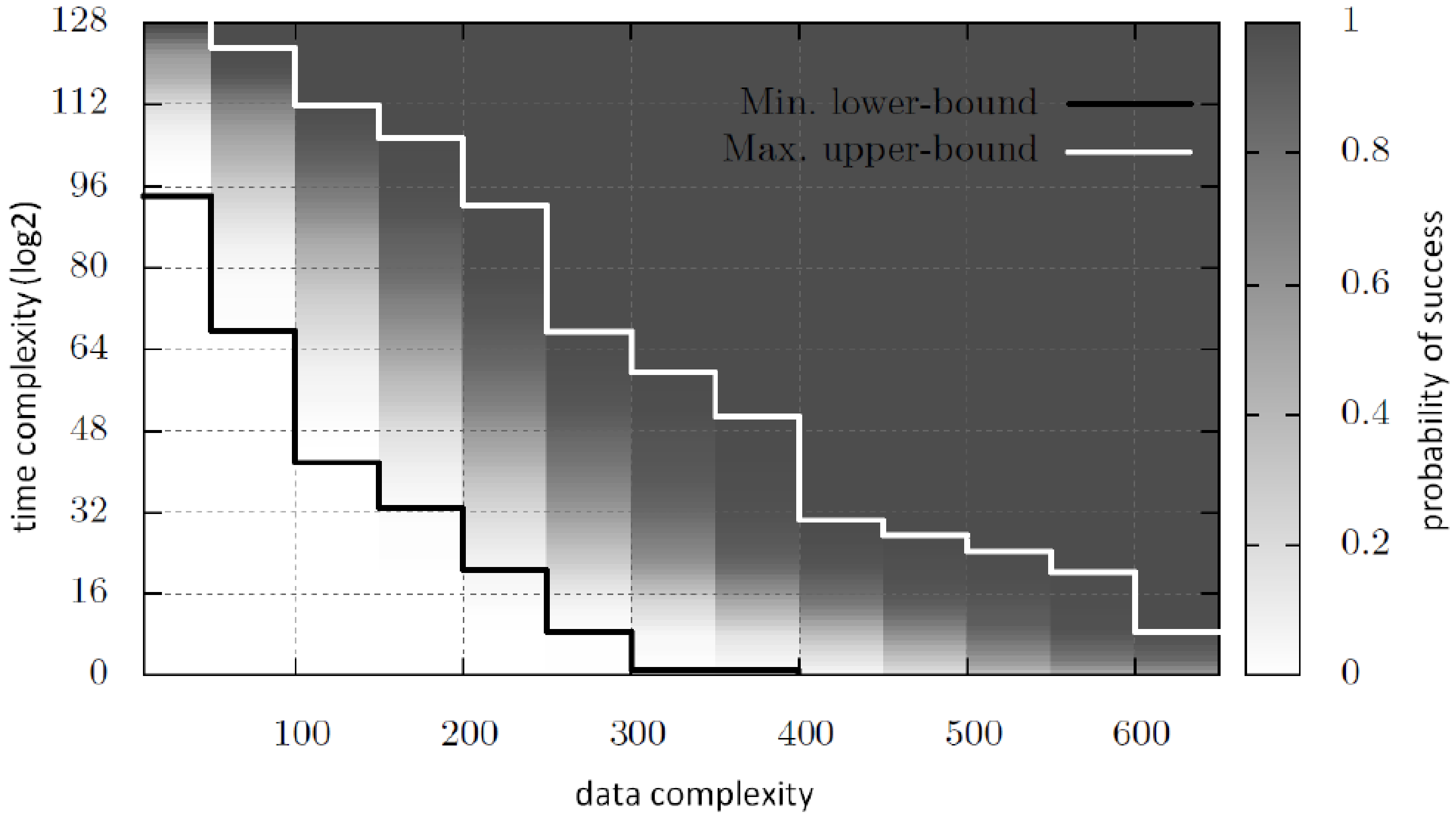
- ***For a fixed number of measurements***
  (which is the case of all real-world evaluations)
  - *If assumption errors are detected*: the loss of information due to an imprecise model is significant (i.e. the model can be improved)
  - *If assumption errors are not detected*: improving the model would not lead to better information extraction (since this improvement could not be distinguished due to the estimation errors)
- All bets are of if more measurements are taken…

Note: also applies to security metrics [3]

- Given a leakage model, it is pretty straightforward to compute security metrics (success probability)

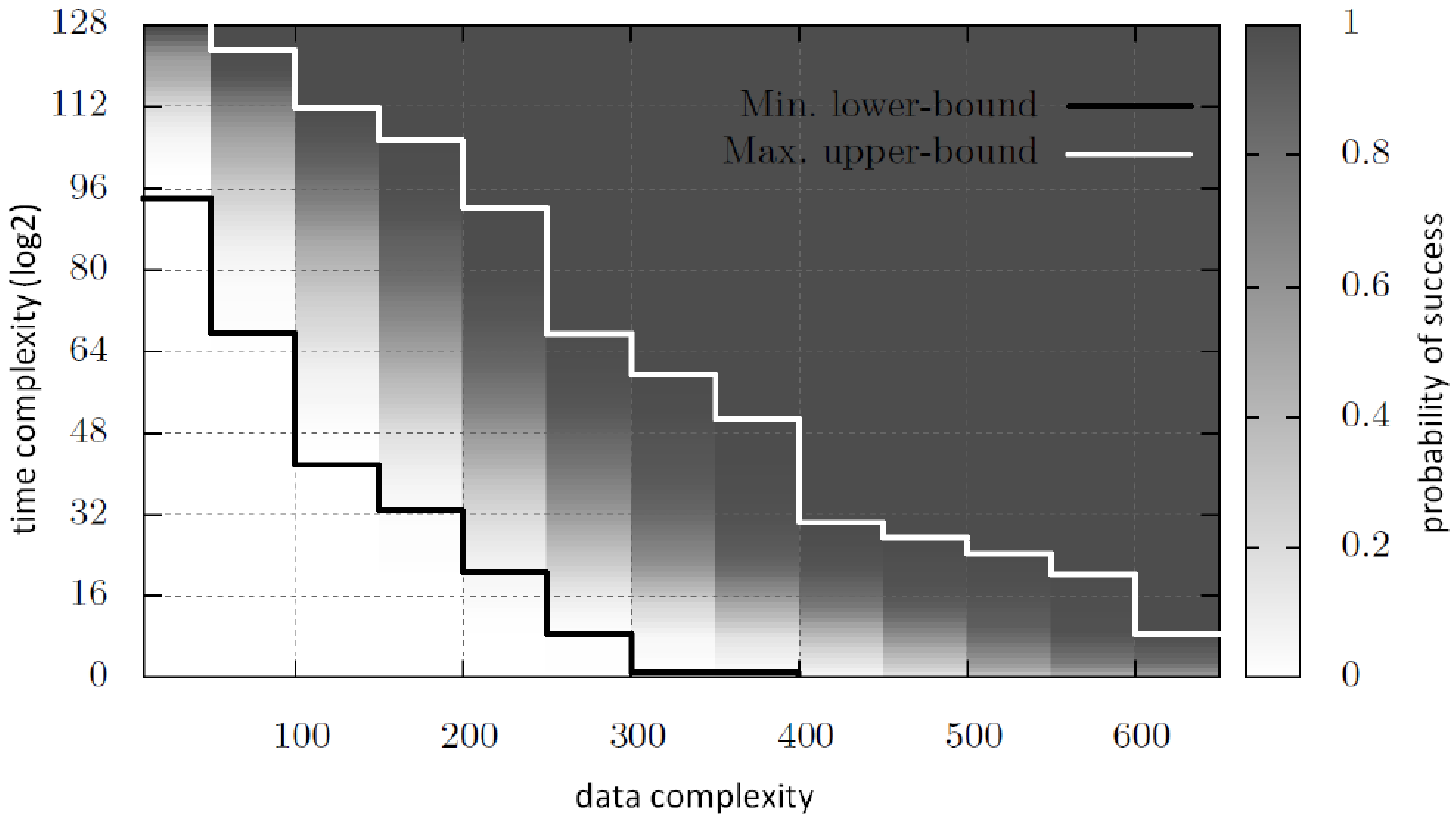
Note: also applies to security metrics [3]

- Given a leakage model, it is pretty straightforward to compute security metrics (success probability)



Note: also applies to security metrics [3]

- Given a leakage model, it is pretty straightforward to compute security metrics (success probability)



- Closer to the ε's in proofs of leakage-resilience

Conclusions

Main message:

- Strict bounds on the information leakage are hard to obtain in general (independent of the distributions and number of measurements)
- But given a number of measurements, we can be sure that a model is “good enough” (or not)

Conclusions

Main message:

- Strict bounds on the information leakage are hard to obtain in general (independent of the distributions and number of measurements)
- But given a number of measurements, we can be sure that a model is "good enough" (or not)
- Quite general problem (not limited to side-channel attacks): applies to any attempt to model an unknown physical or biological process

THANKS

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

1. F.-X. Standaert, T.G. Malkin, M. Yung, *A Unified Framework for the Analysis of Side-Channel Key Recovery Attacks*, in the proceedings of Eurocrypt 2009, Lecture Notes in Computer Science, vol 5479, pp 443-461, Cologne, Germany, April 2009, Springer.

2. M. Renauld, F.-X. Standaert, N. Veyrat-Charvillon, D. Kamel, D. Flandre, *A Formal Study of Power Variability Issues and Side-Channel Attacks for Nanoscale Devices*, in the proceedings of Eurocrypt 2011, Lecture Notes in Computer Science, vol 6632, pp 109-128, Tallinn, Estonia, May 2011, Springer.

3. N. Veyrat-Charvillon, B. Gerard, F.-X. Standaert, *Security Evaluations Beyond Computing Power: How to Analyze Side-Channel Attacks you Cannot Mount?*, to appear in the proceedings of Eurocrypt 2013, Lecture Notes in Computer Science, vol 7881, pp 126-141, Athens, Greece, May 2013, Springer.