Towards a Methodology for Experimental Evaluation in Low-Power Wireless Networking

Romain Jacob
Usman Raza
Lothar Thiele
Carlo Alberto Boano
Marco Zimmerling

CPS-IoTBench Workshop
April 15, 2019

Includes material from Hanspeter Schmid and Alex Huber

“ We need a benchmark for IoT networking. ”

" We need a ***benchmark*** *for IoT networking. "*

⇔ Comparing performance

*“ We need a **benchmark** for IoT networking. ”*

⇔ Comparing performance

⇒ Repeatable experiments

*“ We need a **benchmark** for IoT networking. ”*

$\Leftrightarrow$ Comparing performance

$\Rightarrow$ Repeatable experiments

$\Rightarrow$ Formalize the experimental methodology

Fact 1	The RF environment affects performance of low-power protocol in unpredictable ways.*
+	
Fact 2	Real RF environment cannot be controlled.
=	
Consequence	Performance variability is expected.

* Either unfeasible or unpractical to model

ETH zürich

How do you handle variability?

“ Repeat your experiment. ”

Easy, right?

How do you handle variability?

“ Repeat your experiment. ” Easy, right?

Killer questions

How long should be your experiment?
How many times shoud you repeat it?

How do you handle variability?

“ Repeat your experiment. ” Easy, right?

How long should be your experiment?
How many times shoud you repeat it?

“ Run many long tests. ” Not so easy...

ETH zürich

Let us assume

You ran “ many long tests ”

1 M samples of XYZ
spread over a
large period of time

How do you synthesize your results?

“ Use statistics. ”

ETH zürich

Let us assume	You ran " many long tests "	1 M samples of XYZ spread over a large period of time

How do you synthesize your results?

" Use statistics. "

Literally	A piece of data obtained from a large quantity of data	Mean, median, standard deviation, etc.

ETH zürich

Beware!

Descriptive statistics	$\neq$	Predictive statistics
What is the collected data like		What the collected data allows to infer about future/other data (unknown)

Beware!

Descriptive statistics ≠ Predictive statistics

What is the collected data like

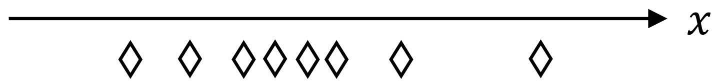
What the collected data allows to infer about future/other data (unknown)

The " interesting case "

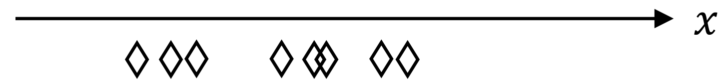
ETH zürich

Descriptive statistics compare the samples

Sample A



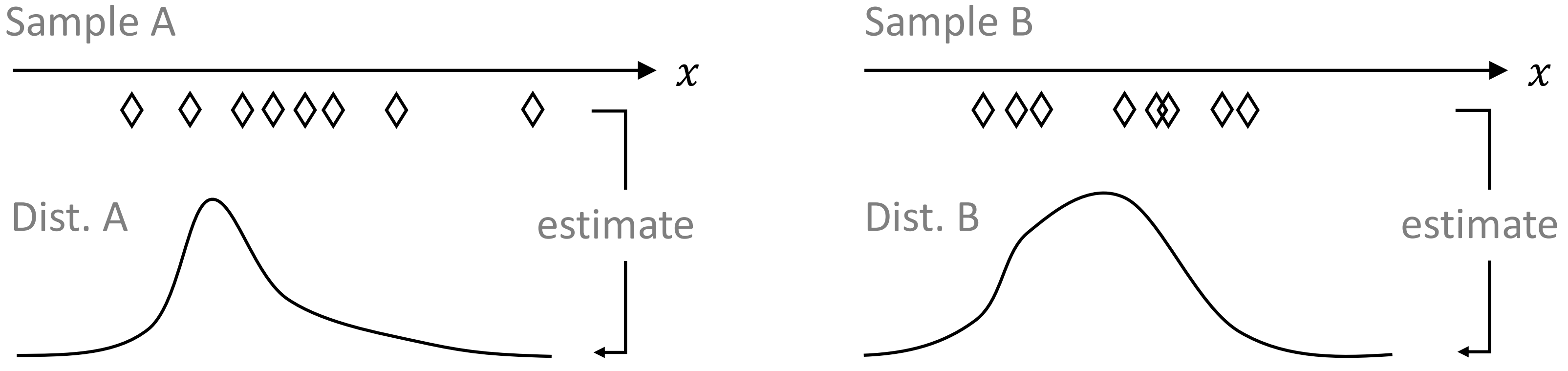
Sample B



Conclusion Sample A is “ better ” than sample B.

ETH zürich

Predictive statistics (aims to) compare the underlying distributions

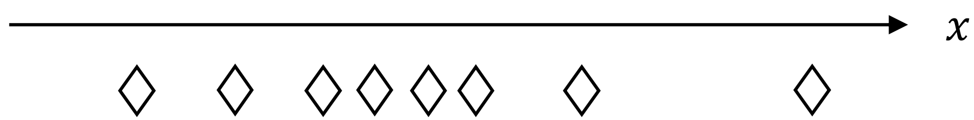


Conclusion	A is " better " than B.	If one sample A and B, then likely sample A is " better " than sample B.

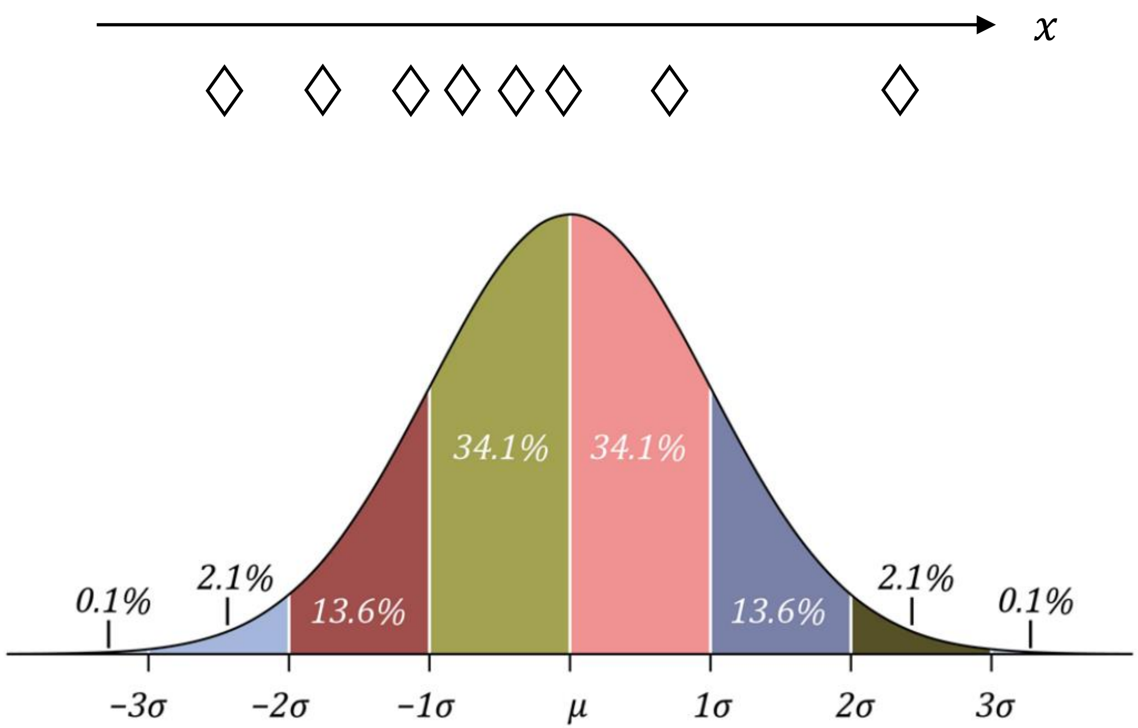
Descriptive statistics	$\neq$	Predictive statistics
Sample A is “ better ” than sample B.		If one sample A and B, then likely sample A is “ better ” than sample B.

Much stronger statement
but also harder to make

ETH zürich



Tendency	Mean	Prediction?
Variability	Standard deviation	



ETH zürich

If we use mean and standard deviation in this context, we make two mistakes

First error

The standard deviation of the sample is not
the standard deviation of the underlying distribution.

Use confidence intervals

Confidence interval (CI) ?

Informally

A numerical interval
in which lies the true value (which you don't known)
of some parameter with some probability (or confidence level)

Example

$[a, b]$ is a 95% CI for the mean of x

$\Leftrightarrow$ The probability than the true mean value of x is included in $[a, b]$ is larger or equal to 95%

If we use mean and standard deviation in this context, we make two mistakes

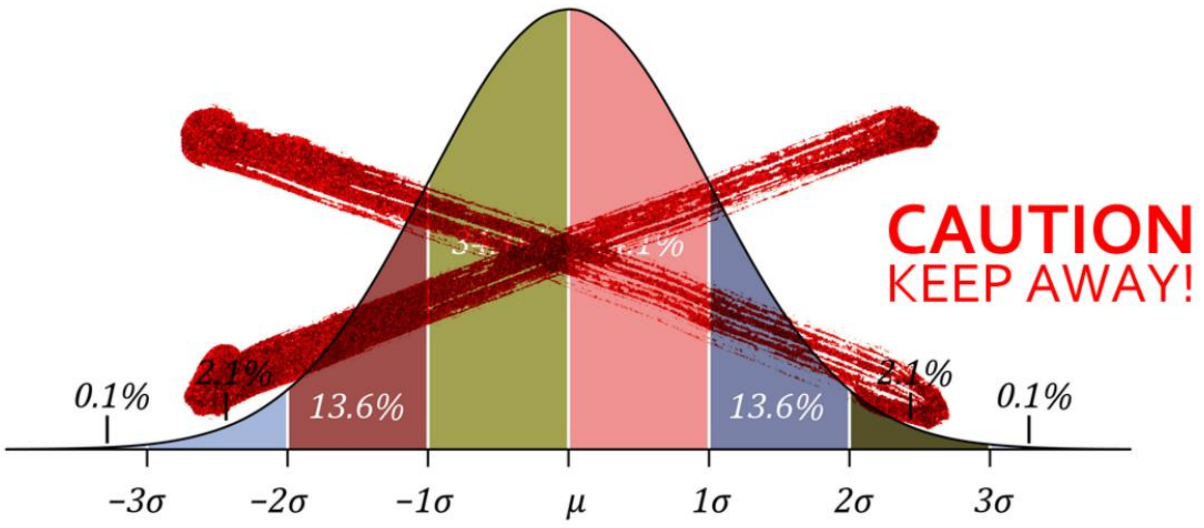
First error — The standard deviation of the sample is not the standard deviation of the underlying distribution.

Use confidence intervals

Second error — The underlying distribution is not normal!

The standard deviation does not help making predictions
Use non-parametric statistics

ETH zürich

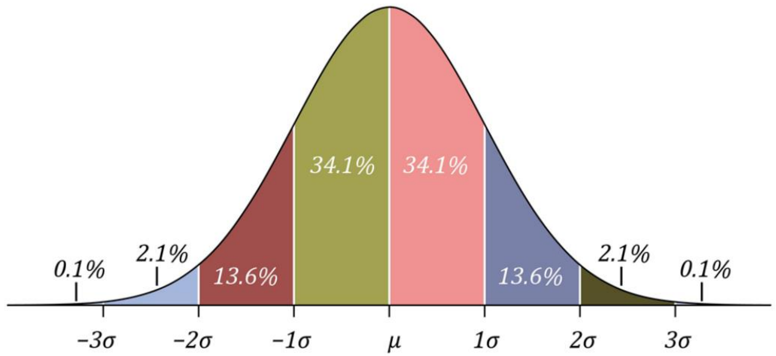

CAUTION
KEEP AWAY!
0.1%
2.1%
13.6%
13.6%
2.1%
0.1%
−3σ
−2σ
−1σ
μ
1σ
2σ
3σ

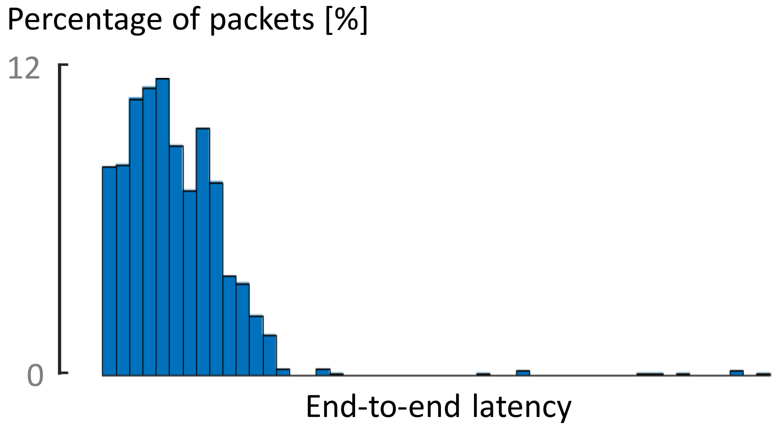
Towards a Methodology for Experimental Evaluation in Low-Power Wireless Networking

Know your data
Use non-parametric statistics

Formalizing low-power wireless experimental evaluation

Performance measurements in computer science are typically not normally distributed





ETH zürich

Performance measurements in computer science are typically not normally distributed

Use non-parametric statistics based on distribution percentiles — Great for predictive statistics

p-th percentile or Pp	$p\%$	of the distribution is below
	$(1-p)\%$	of the distribution is above

Percentiles are powerful predictive statistics

Simple to use	Can compute CI for any percentile with any confidence
Distribution independent	Estimates are valid regardless of the underlying distribution
Robust	Estimates are not skewed by outliers

Confidence intervals

$$\mathsf{P}\left\{x_m \leq M \leq x_{N-m+1}\right\} = 1 - 2\sum_{k=0}^{m-1}\binom{N}{k}\frac{1}{2^N}$$

$$\mathsf{P}\left\{x_m \leq P_p\right\} =$$

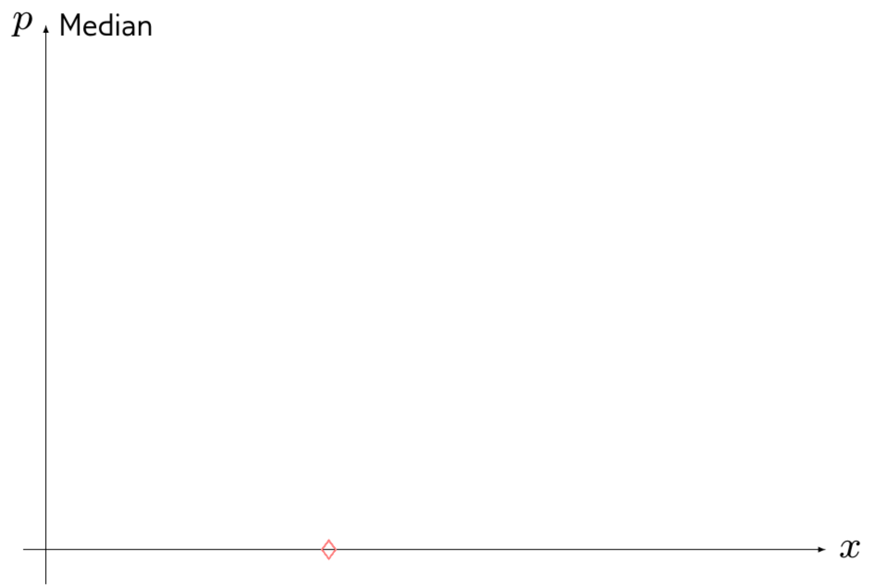
$$P\left\{x_{N-m+1} \geq P_{1-p}\right\} = 1 - \sum_{k=0}^{m-1}\binom{N}{k}p^k(1-p)^{N-k}$$

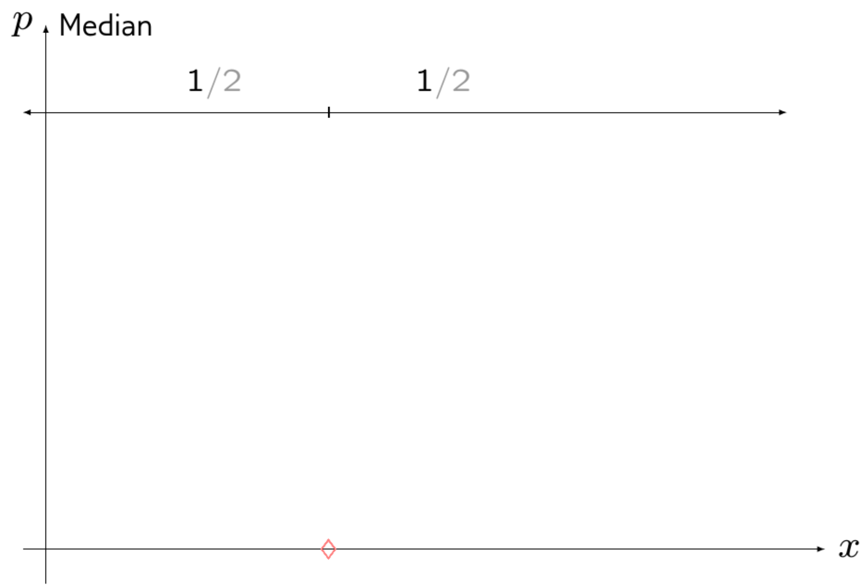
Confidence intervals

$$\mathsf{P}\left\{x_m \leq M \leq x_{N-m+1}\right\} = 1 - 2\sum_{k=0}^{m-1}\binom{N}{k}\frac{1}{2^N}$$

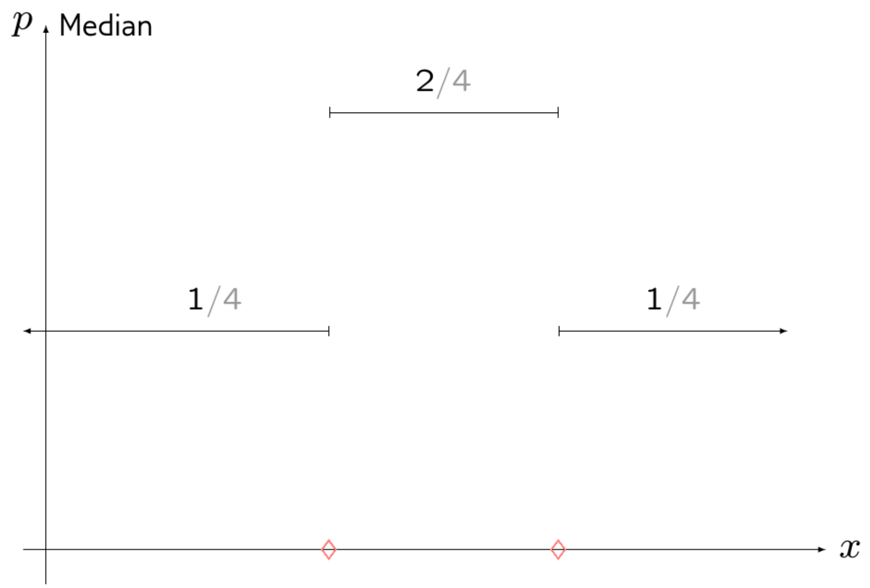
$$\mathsf{P}\left\{x_m \leq P_p\right\} =$$

$$P\left\{x_{N-m+1} \geq P_{1-p}\right\} = 1 - \sum_{k=0}^{m-1}\binom{N}{k}p^k(1-p)^{N-k}$$


p
Median
x

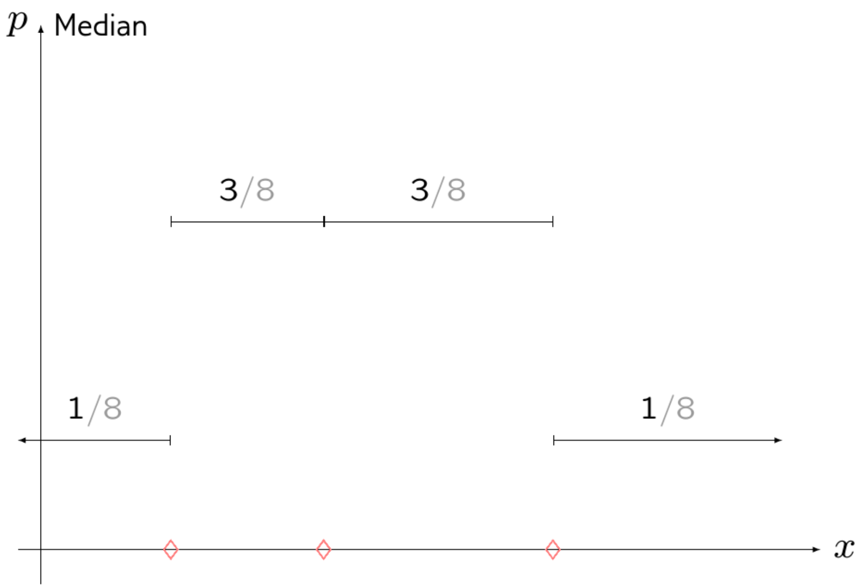

Median
p
1/2
1/2
x

Hypothesis Samples are i.i.d.



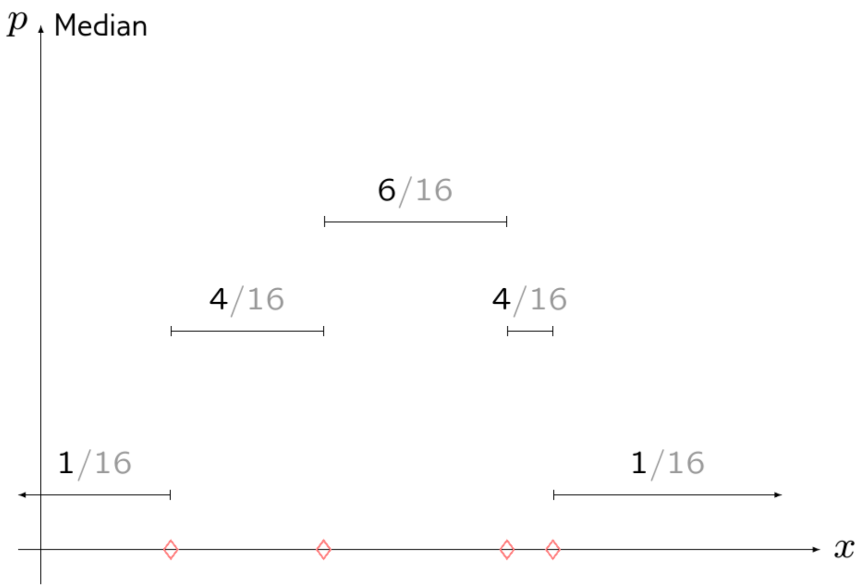
ETH zürich

Hypothesis Samples are i.i.d.



ETH zürich

Hypothesis Samples are i.i.d.



ETH zürich

Binomial distribution

$$\mathsf{P}\left\{x_k \leq P_p \leq x_{k+1}\right\} = \binom{N}{k} p^k (1-p)^{N-k}$$

ETH zürich

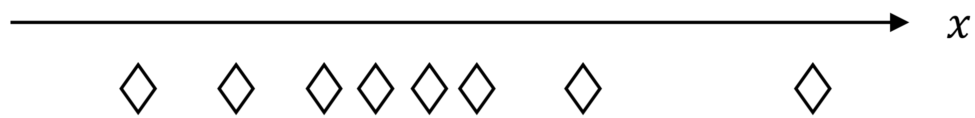
Confidence intervals

$$\mathsf{P}\left\{x_m \leq M \leq x_{N-m+1}\right\} = 1 - 2\sum_{k=0}^{m-1}\binom{N}{k}\frac{1}{2^N}$$

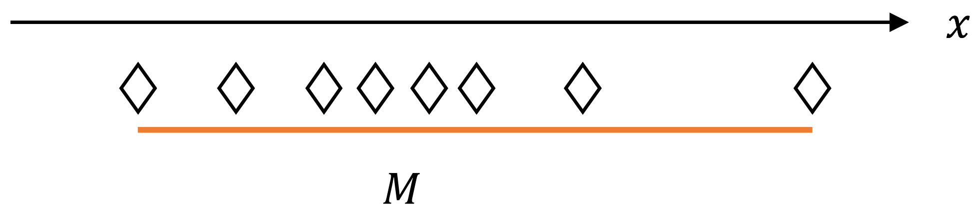
$$\mathsf{P}\left\{x_m \leq P_p\right\} =$$

$$P\left\{x_{N-m+1} \geq P_{1-p}\right\} = 1 - \sum_{k=0}^{m-1}\binom{N}{k}p^k(1-p)^{N-k}$$

ETH zürich



ETH zürich



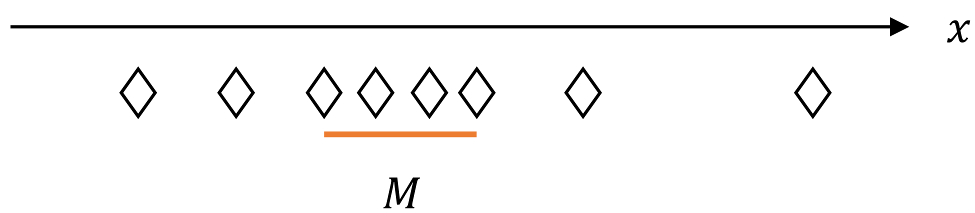
$$P = 99{,}29\ \%$$

when 2 x 0 points are excluded

ETH zürich

$$P = 92{,}97\ \%$$

when 2 x 1 points are excluded



$$P = 71{,}09\ \%$$

when 2 x 2 points are excluded

ETH zürich

$$P = 27{,}34\ \%$$

when 2 x 3 points are excluded

ETH zürich

Percentiles are powerful predictive statistics

Simple to use	Can compute CI for any percentile with any confidence
Distribution independent	Estimates are valid regardless of the underlying distribution
Robust	Estimates are not skewed by outliers

Percentiles are powerful predictive statistics

Simple to use	Can compute CI for any percentile with any confidence
Distribution independent	Estimates are valid regardless of the underlying distribution
Robust	Estimates are not skewed by outliers

For any confidence c, any percentile p,

$$N \geq \frac{\log(1-c)}{\log(1-p)}$$

Example

$c = 0{,}95$

$p = 0{,}01$ or 1-th percentile

$$\Rightarrow N \geq 299$$

ETH zürich

Thus

We can derive the minimal number of samples required for estimating any percentile with any confidence

Thus

We can derive the minimal number of samples required for estimating any percentile with any confidence

So now

How long should be your experiment?
How many times shoud you repeat it?

Can be answered rationally.

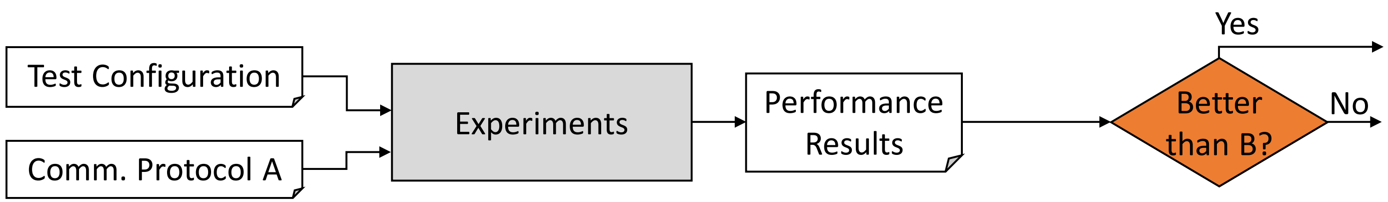
ETH zürich

Towards a Methodology for Experimental Evaluation in Low-Power Wireless Networking

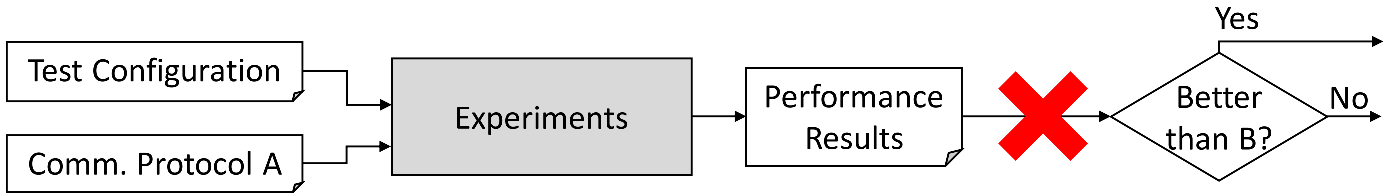
Know your data

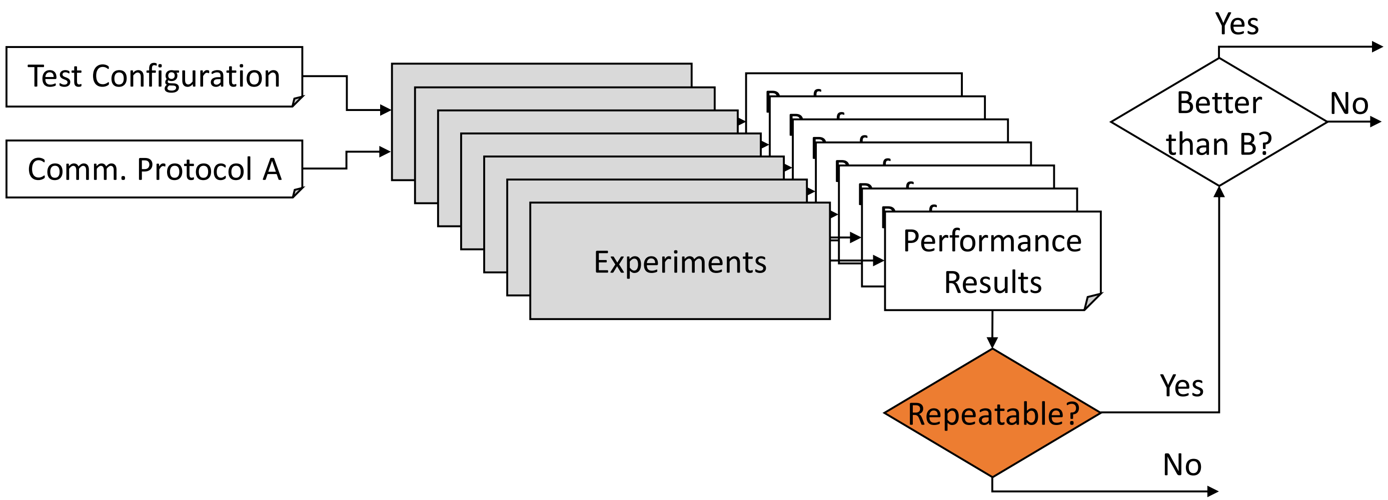
Use non-parametric statistics

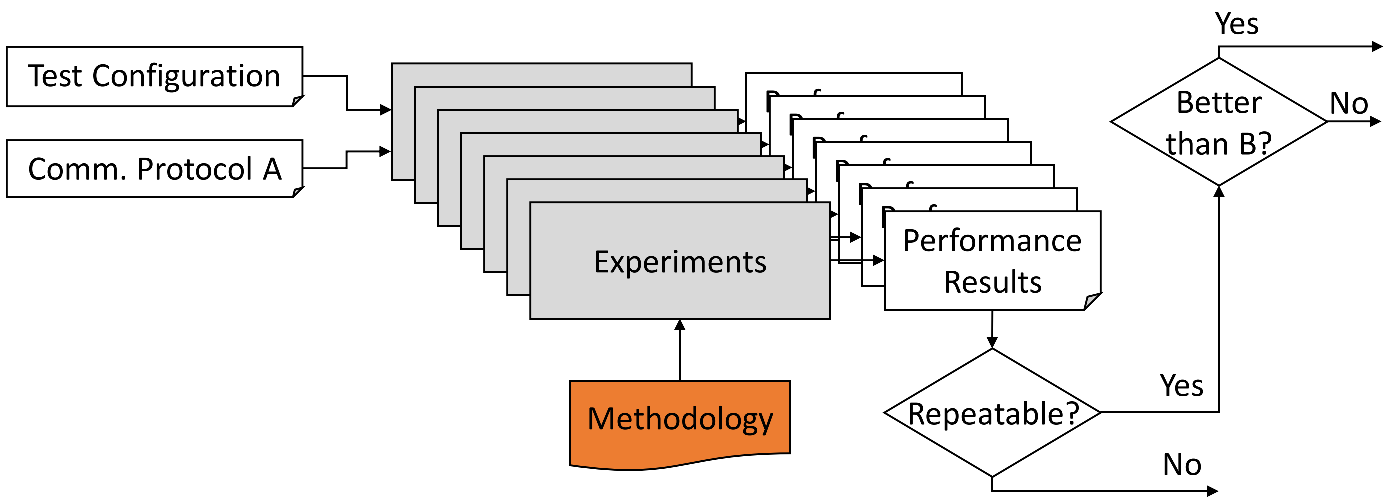
Formalizing low-power wireless experimental evaluation

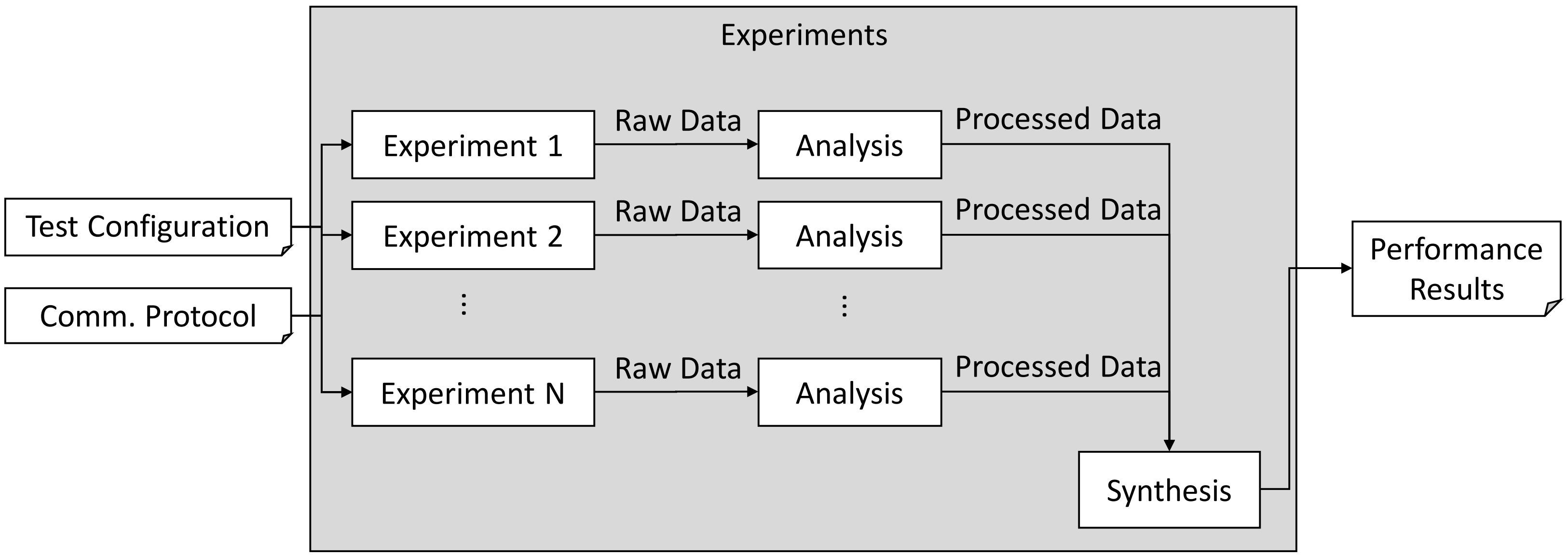

Test Configuration
Comm. Protocol A
Experiments
Performance Results
Better than B?
Yes
No

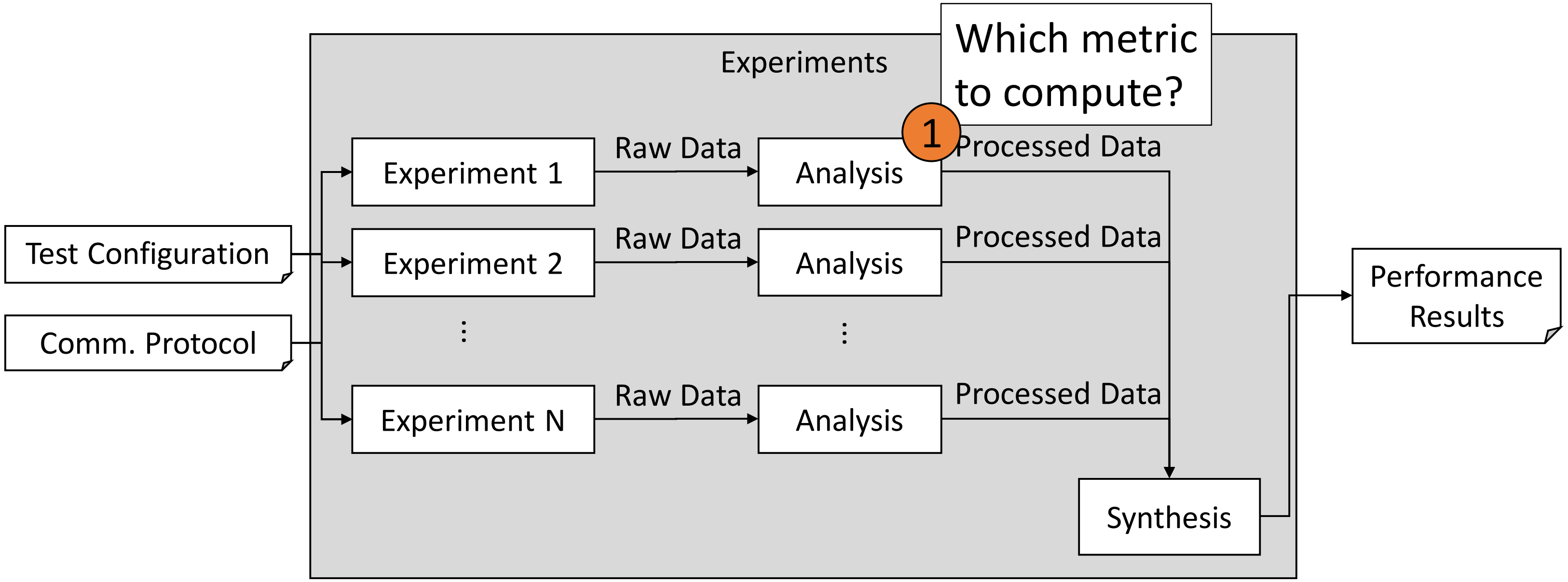
ETH zürich


Test Configuration
Comm. Protocol A
Experiments
Performance Results
Better than B?
Yes
No

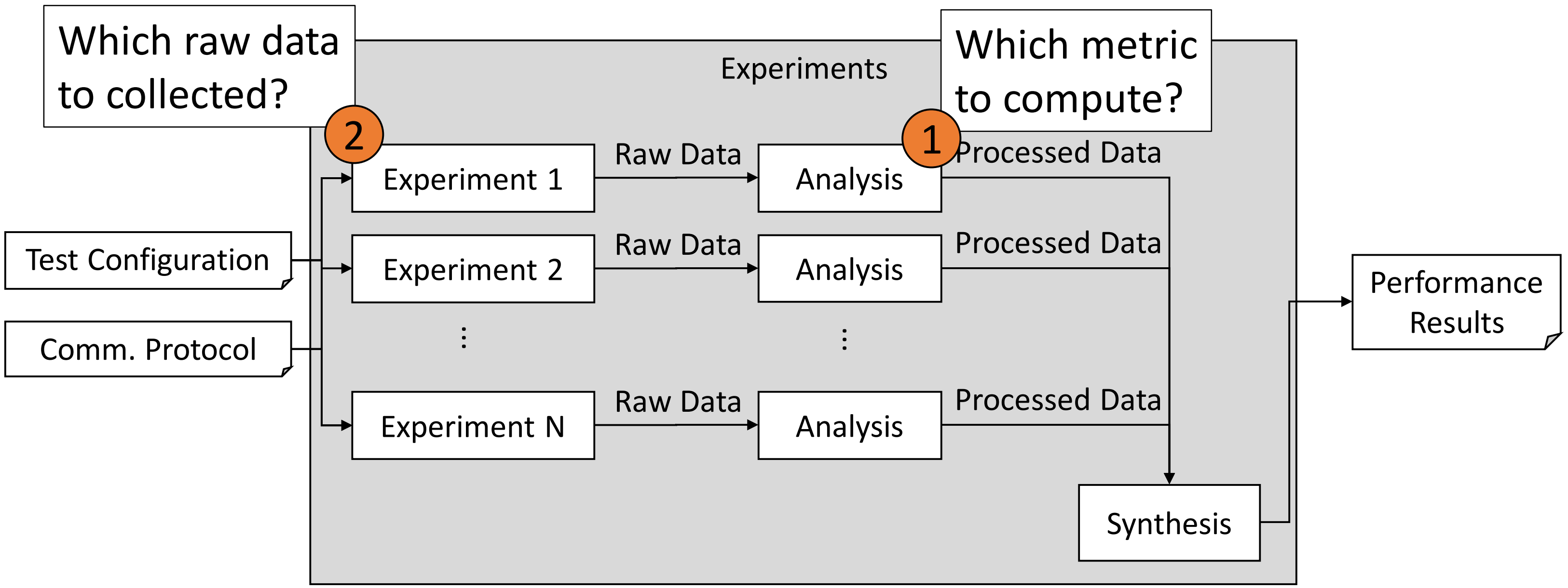

Test Configuration
Comm. Protocol A
Experiments
Performance Results
Repeatable?
Yes
No
Better than B?
Yes
No

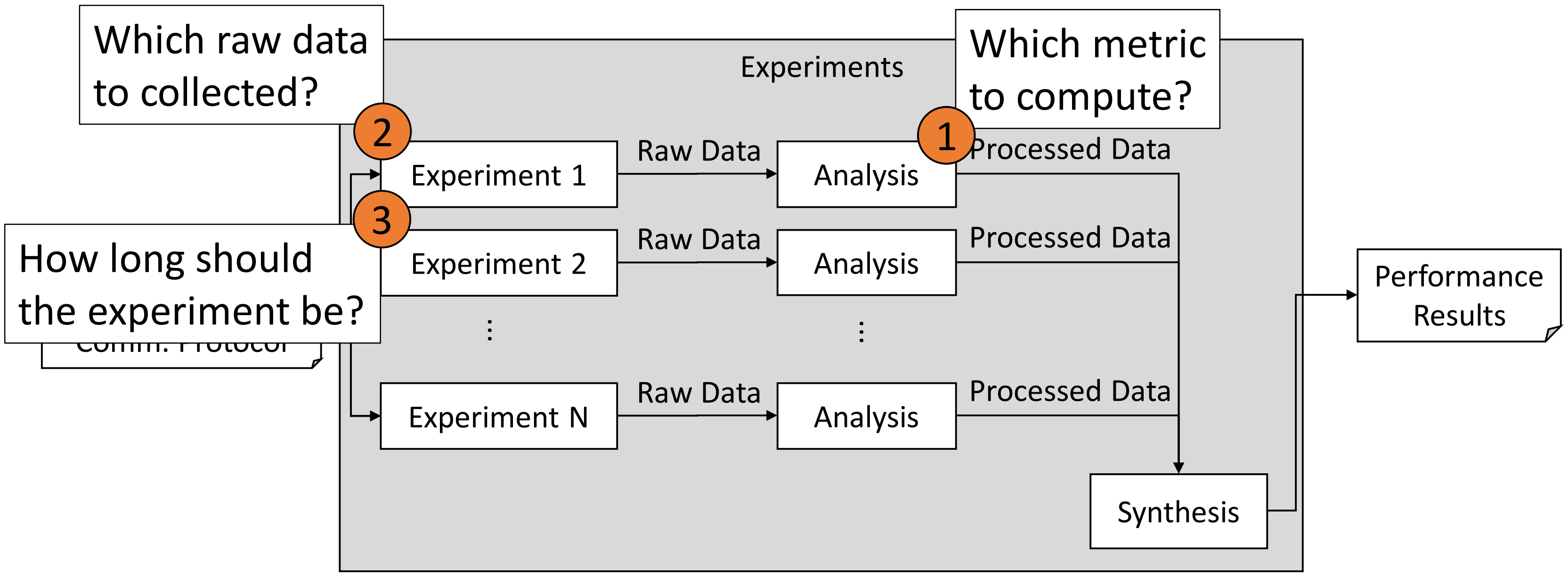

Test Configuration
Comm. Protocol A
Experiments
Methodology
Performance Results
Repeatable?
Yes
No
Better than B?
Yes
No

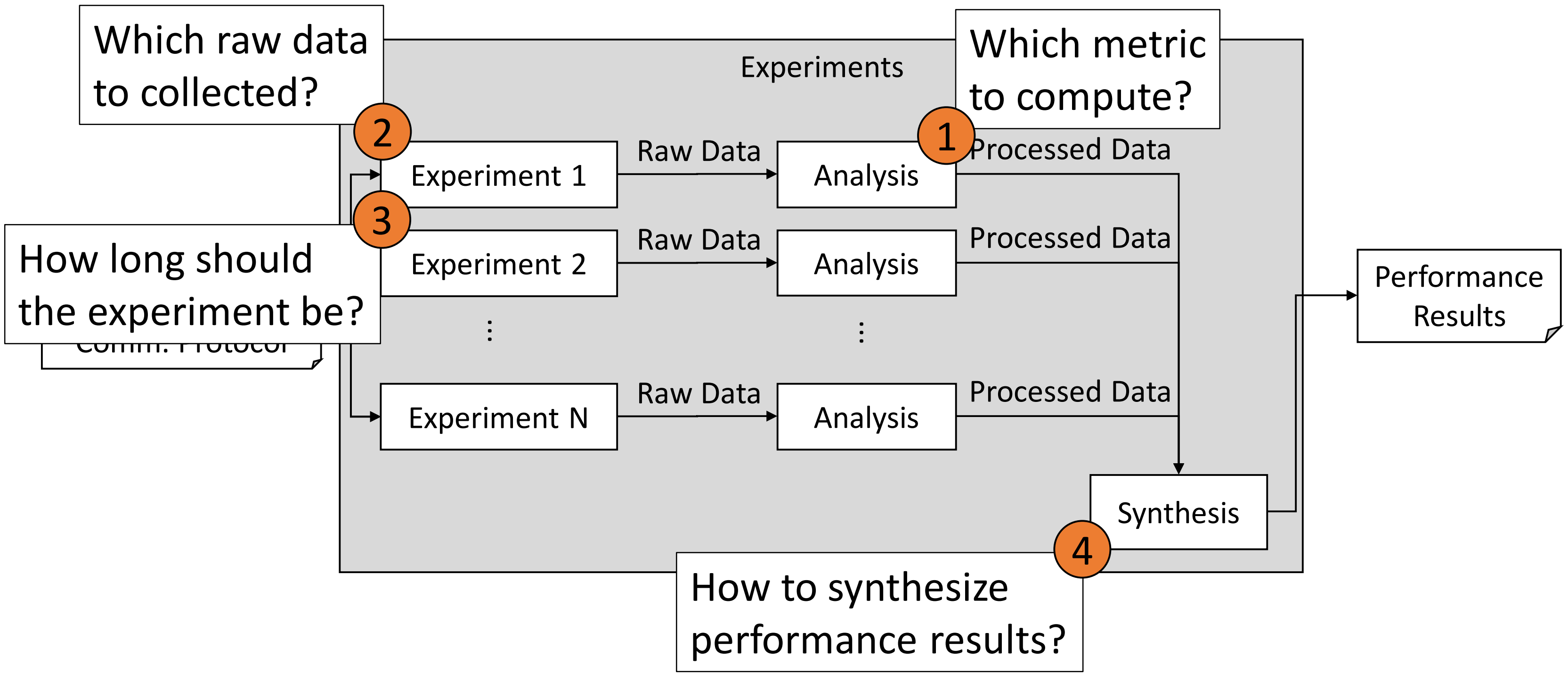

Experiments
Test Configuration
Comm. Protocol
Experiment 1
Experiment 2
Experiment N
Raw Data
Raw Data
Raw Data
Analysis
Analysis
Analysis
Processed Data
Processed Data
Processed Data
Synthesis
Performance Results

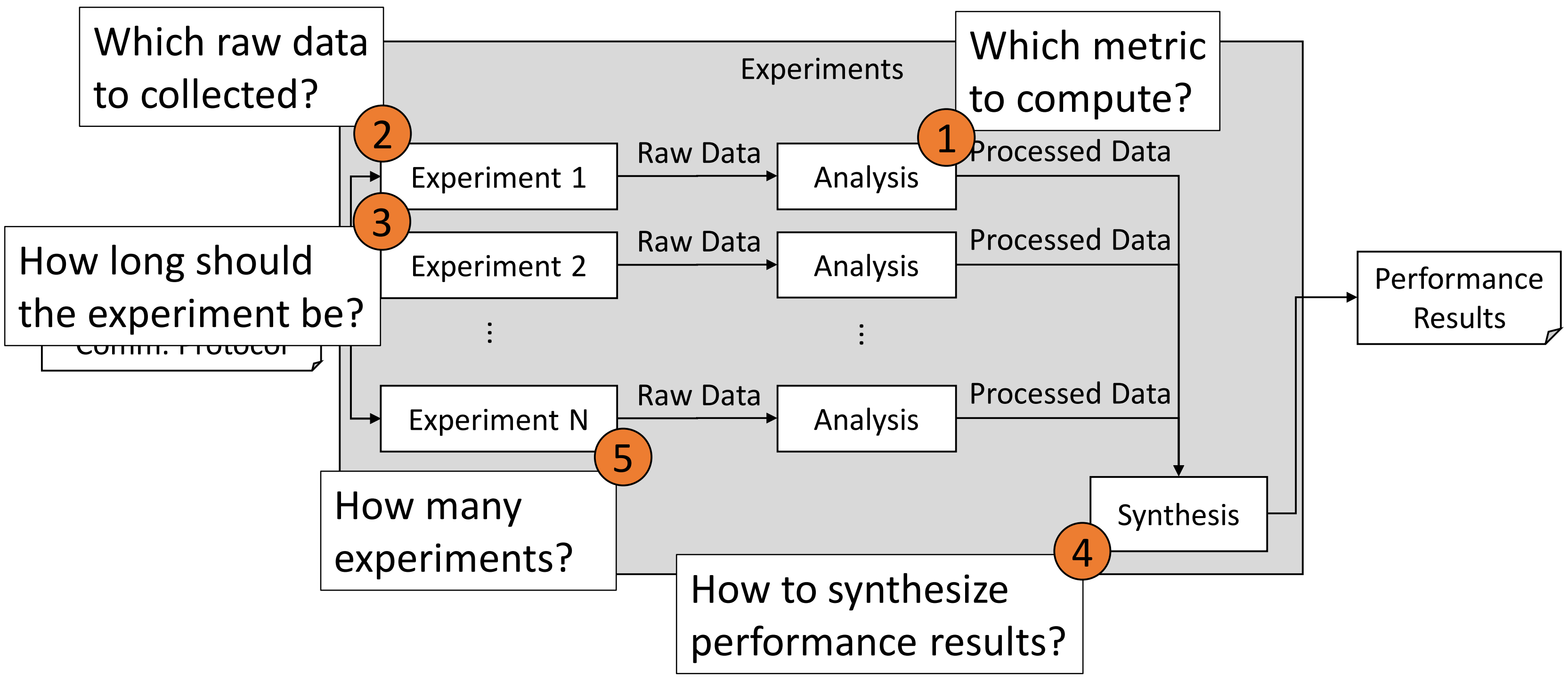

Which metric to compute?
Experiments
1
Test Configuration
Comm. Protocol
Experiment 1
Experiment 2
Experiment N
Raw Data
Raw Data
Raw Data
Analysis
Analysis
Analysis
Processed Data
Processed Data
Processed Data
Synthesis
Performance Results

ETH zürich


Which raw data to collected?
Which metric to compute?
Experiments
2
1
Test Configuration
Comm. Protocol
Experiment 1
Experiment 2
⋮
Experiment N
Raw Data
Raw Data
Raw Data
Analysis
Analysis
⋮
Analysis
Processed Data
Processed Data
Processed Data
Synthesis
Performance Results


Which raw data to collected?
Which metric to compute?
How long should the experiment be?
Experiments
1
2
3
Experiment 1
Experiment 2
Experiment N
Raw Data
Raw Data
Raw Data
Analysis
Analysis
Analysis
Processed Data
Processed Data
Processed Data
Synthesis
Performance Results


Which raw data to collected?
Experiments
Which metric to compute?
2
Experiment 1
Raw Data
Analysis
1
Processed Data
3
How long should the experiment be?
Experiment 2
Raw Data
Analysis
Processed Data
Performance Results
Experiment N
Raw Data
Analysis
Processed Data
Synthesis
4
How to synthesize performance results?

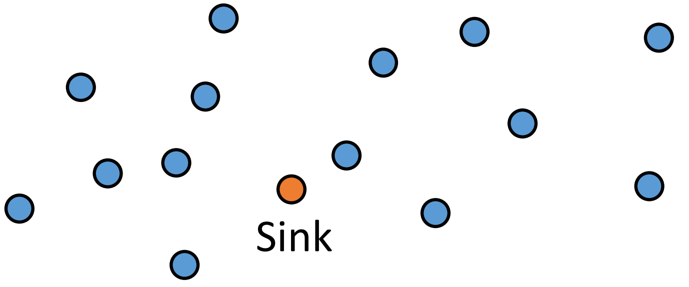

Which raw data to collected?
Experiments
Which metric to compute?
2
Experiment 1
Raw Data
Analysis
1
Processed Data
3
How long should the experiment be?
Comm. Protocol
Experiment 2
Raw Data
Analysis
Processed Data
Performance Results
Experiment N
Raw Data
Analysis
Processed Data
5
How many experiments?
Synthesis
4
How to synthesize performance results?

ETH zürich

Case study – Periodic data collection

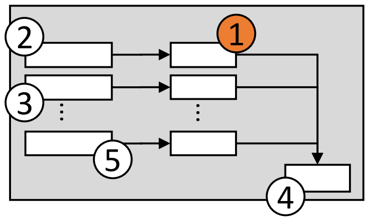
14	source nodes
200	payloads per source
2	Bytes per payload
10	payload per second
	Periodic release, asynchronous

First payload released after 10s
Test stops 10s after last payload is released



ETH zürich

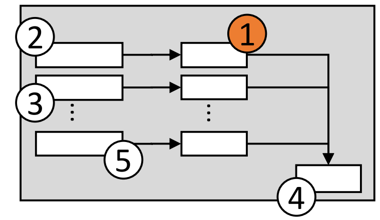
Select the "metrics" based on the purpose of the evaluation



Performance dimensions	Reliability Latency Energy efficiency	Average Extremal
Metrics	Radio duty-cycle Current draw	
Measures	Mean Median/percentiles Max/Min	

ETH zürich

Select the "metrics" based on the purpose of the evaluation



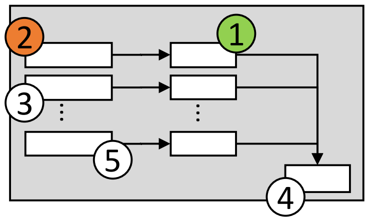
How many application payloads can one expect to successfully receive in one execution of the scenario?

Dimension	Metric	Measure
Average reliability	PRR	~~Mean~~ Median

We are trying to predict future performance

ETH zürich

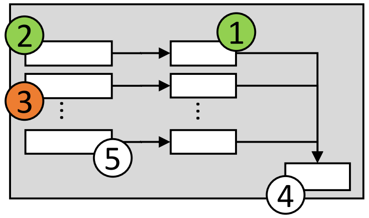
Collect raw data with the finest granularity possible



PRR	Log all received payloads at the sink	
Current draw	14400 samples/s 10 pA precision	1 every ~7μs

ETH zürich

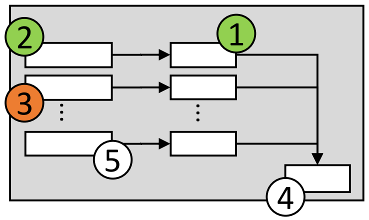
Define the length of the experiment based on the scenario and the protocol



Generally difficult	Correct approach depends on the protocol under test
Easy case	If scenario is terminating and short, then run it in full

ETH zürich

Define the length of the experiment based on the scenario and the protocol



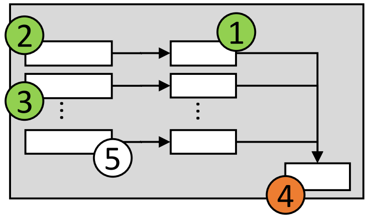
	Terminating	200 payload/source
+	Short	10 + 200/10 + 10 = 40s

⇒ Run in full

⇒ The uncertainty lies only in the variability across experiments

ETH zürich

Use performance indicators based on confidence intervals



Average reliability

Overall PRR

$\Rightarrow \forall$ experiment j, x_j

$$\frac{Received\ payloads}{200 * 14}$$

95% CI on the median PRR for all exp.

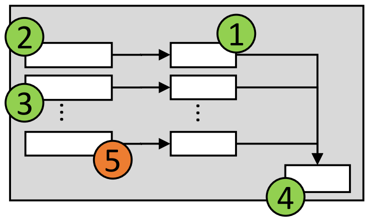
$\Rightarrow [x_m, x_{N-m+1}]$

Use conservative bound

$\Rightarrow x_m$

ETH zürich

Perform sufficiently many experiments to obtain tight CI

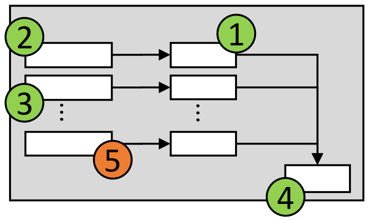


Intuition

Estimating	average	performance is easier
than	extremal	performance

Need more experiments to estimate a 95-th percentile than a median

Perform sufficiently many experiments to obtain tight CI



Performance indicators based on 95% CI for the median

$\Rightarrow N \geq 6$

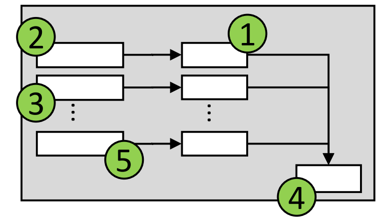
Aim for tigher CI

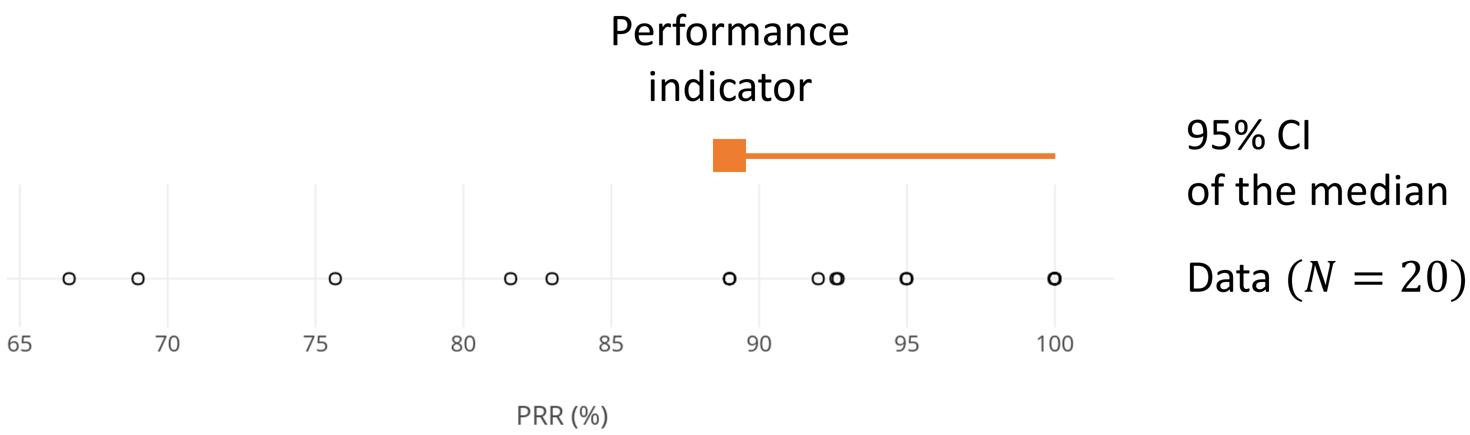
$N = 20$

$\Rightarrow$ 95% CI is $[x_6, x_{15}]$

ETH zürich

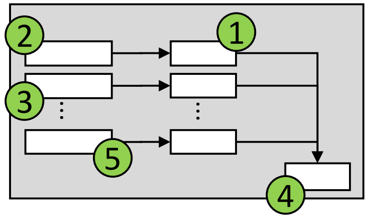
Following the methodology enables unambiguous performance reports





ETH zürich

Following the methodology enables unambiguous performance reports



Protocol	A	B	C	D	E	F	G
Average Energy	0.82	0.83	0.89	0.86	0.90	0.43	0.25
Worst-case Energy	0.67	0.44	0.82	0.18	0.52	0.27	0.19
Reliability	0.40	0.41	0.89	0.06	0.48	0.27	0.25

We are on good way...

ETH zürich

*" We need a **benchmark** for IoT networking. "*

$\Leftrightarrow$ Comparing performance

$\Rightarrow$ Repeatable experiments

$\Rightarrow$ Formalize the experimental methodology

2nd Workshop on Benchmarking Cyber-Physical Networks and Internet of Things (CPS-IoTBench)
April 15, 2019 | Montréal | Canada

Towards a Methodology for Experimental Evaluation in Low-Power Wireless Networking

Know your data
Use non-parametric statistics

Formalizing low-power wireless
experimental evaluation

Problem 1	Predictive statistics require i.i.d. measurements Not a given. This must be checked, not assumed.
Problem 2	What if the scenario is not terminating? We still don't know how long one experiement should be.
Problem 3	To be comparable, results must be repeatable We still don't know how to formalize repeatability in our context.

These are work-in-progress...

ETH zürich

2nd Workshop on Benchmarking Cyber-Physical Networks and Internet of Things (CPS-IoTBench)
April 15, 2019 | Montréal | Canada

Towards a Methodology for Experimental Evaluation in Low-Power Wireless Networking

Romain Jacob
Usman Raza
Lothar Thiele
Carlo Alberto Boano
Marco Zimmerling

ETH zürich
TOSHIBA
TU Graz
TECHNISCHE UNIVERSITÄT DRESDEN

Includes material from Hanspeter Schmid and Alex Huber