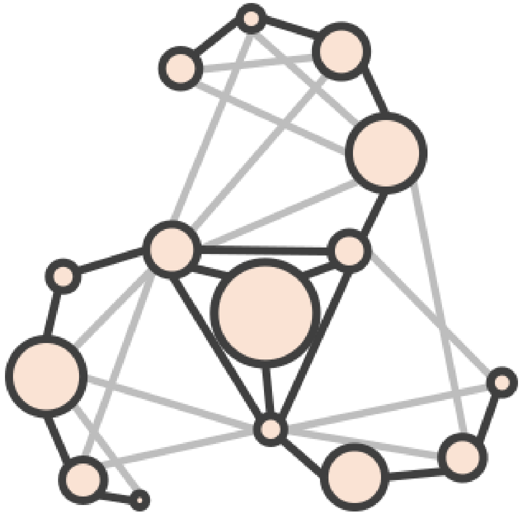
Experimental Reproducibility in Networking Research



Romain Jacob
ETH Zurich

GT Reproducibilité
GDR Réseaux et Systèmes Distribués
May 10, 2022

@RJacobPartner

2015—2019	Doctorate from ETH Zurich with Prof. Lothar Thiele in real-time communication protocols for low-power embedded systems
Since then	PostDoc in Computer Networks with Prof. Laurent Vanbever focus on protocol design for “greening” the Internet

Key questions	1. How to design experiments?	
	2. How to analyse data?	
Focus	Networking	Field
	Performance evaluations	Exp. type
Goal	Foster replicability	
	To be clarified	

This is an interactive session
Questions are welcome!

- Write any question in the chat;
- There will be several time slots for questions

Don't be shy

Direct question by voice
are welcome during the Q&A

Please

Stay muted during
the rest of the presentation

45’ Lecture

10’ Hands-on

10’ Break

20’ Lecture

Wrap-up & Discussions

45’	Lecture
10’	Hands-on
10’	Break
20’	Lecture
	Wrap-up & Discussions

Why replicability matters

Case by example

Understanding variability

The three timescales

Know your data

Use the right statistics

Why replicability matters

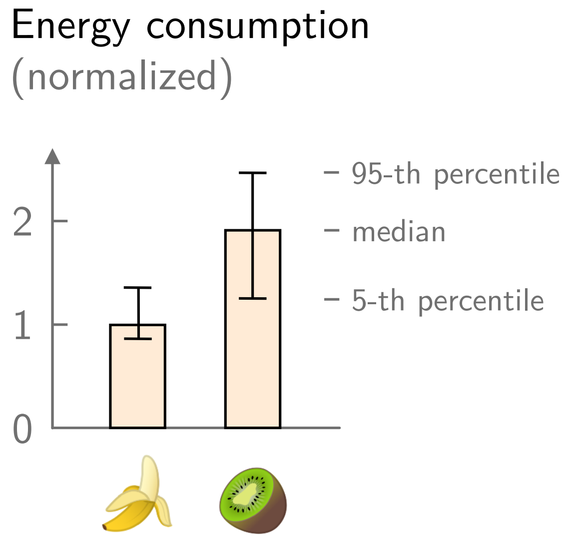
Case by example

Understanding variability

The three timescales

Know your data

Use the right statistics

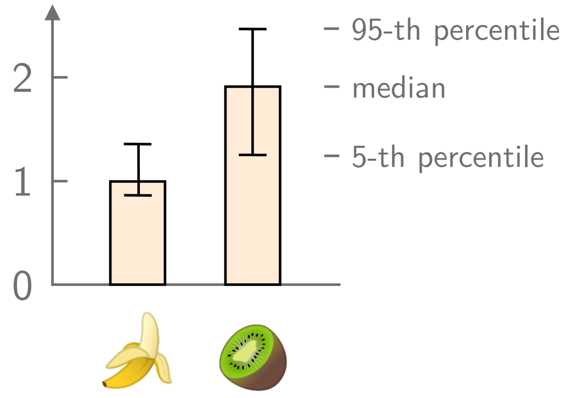


A team designed Banana, a new (and amazing!) ultra-low-power wireless communication protocol.

They set up an experiment to validate their claims.

- They deploy Banana on a real-world testbed;
- They run one benchmark problem for data collection from the IoTBench;
- They compare Banana's performance against the state-of-the-art Kiwi protocol, which is re-run as part of the experiment.
- Each protocol is tested 10 times.

Energy consumption
(normalized)



Claim

“🍌 achieves a 2x improvement over 🥝.”

You are reviewing the paper

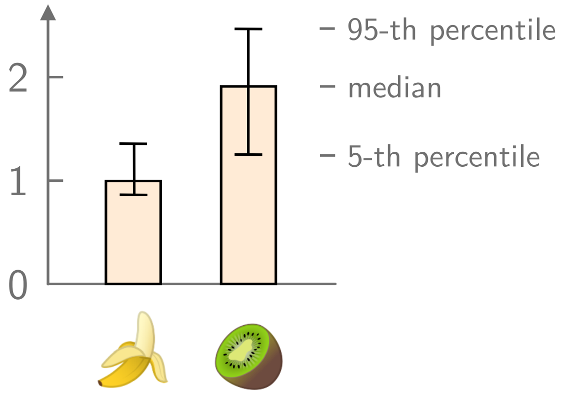
Are ten runs enough to support this claim?

slido

Are ten runs enough?

ⓘ Start presenting to display the poll results on this slide.

Energy consumption
(normalized)



Claim

"🍌 achieves a 2x improvement over 🥝."

You are reviewing the paper

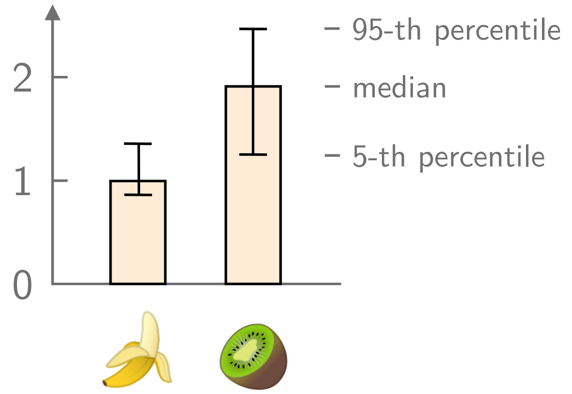
Are ten runs enough to support this claim?

How many runs do you think are required?

How many do you think are required?

ⓘ Start presenting to display the poll results on this slide.

Energy consumption
(normalized)



Claim

"🍌 achieves a 2x improvement over 🥝."

You are reviewing the paper

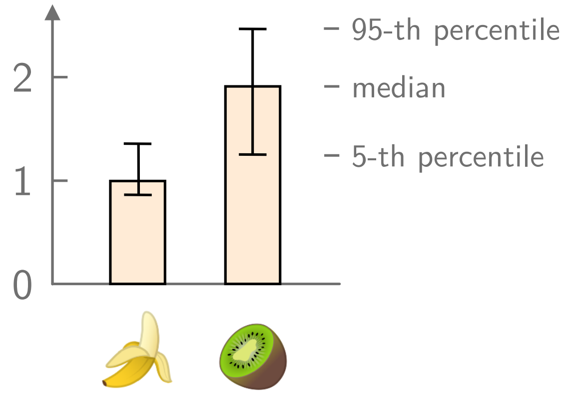
Are ten runs enough to support this claim?

How many runs do you think are required?

Cannot say.

▶ Which "performance" are we talking about?

Energy consumption (normalized)



Claim

"🍌 achieves a 2x improvement over 🥝, in the median case."

You are reviewing the paper

Are ten runs enough to support this claim?

How many runs do you think are required?

Cannot say.

▶ Which "performance" are we talking about?

If you would repeat the experiment, do you think you would obtain the same result?

Hard to say.

▶ What does "same result" mean, really?

These are
hard questions!

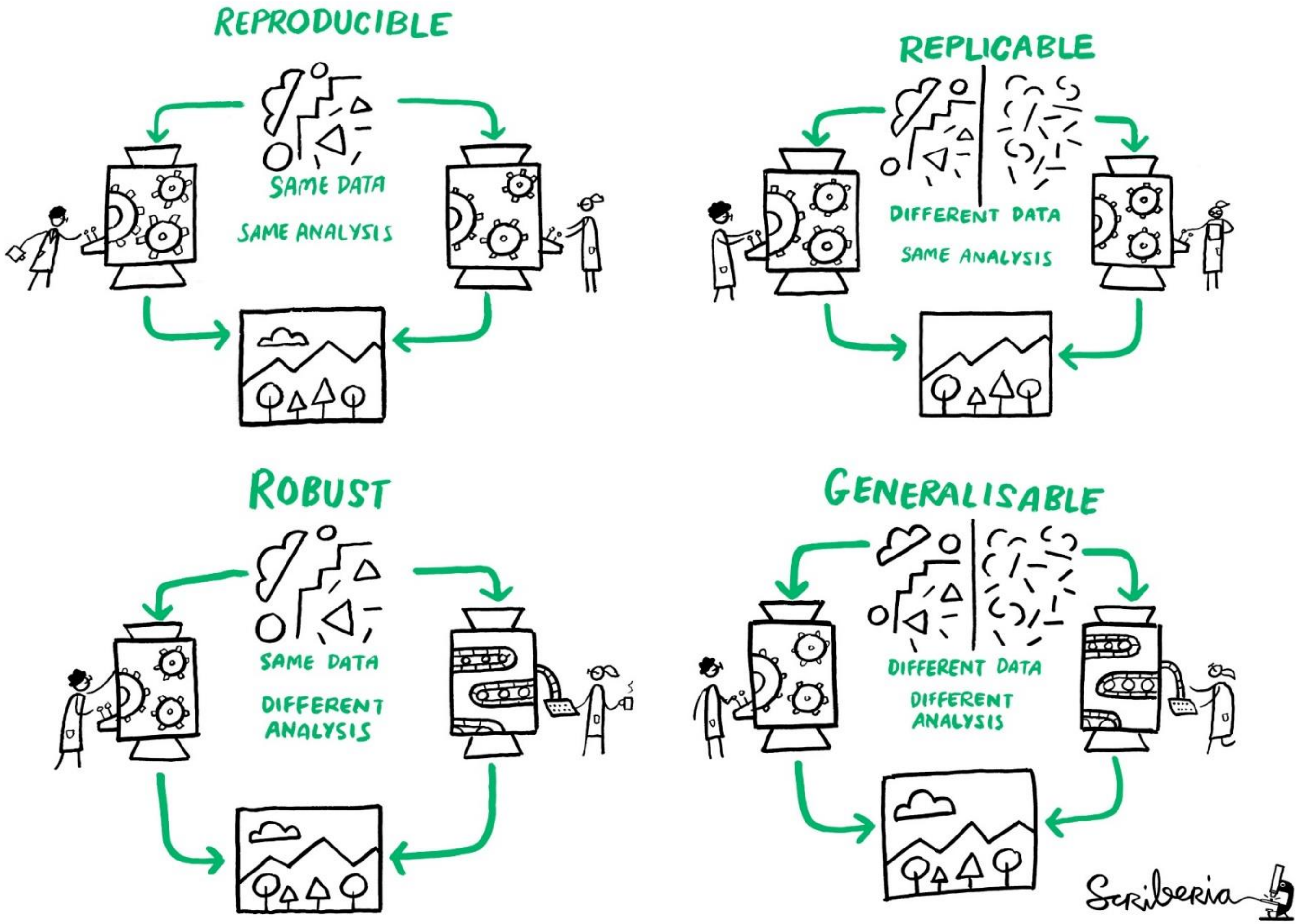
How many runs
are required?

Would you obtain
the same results?

This tutorial presents a rational methodology
to address these questions (and others)

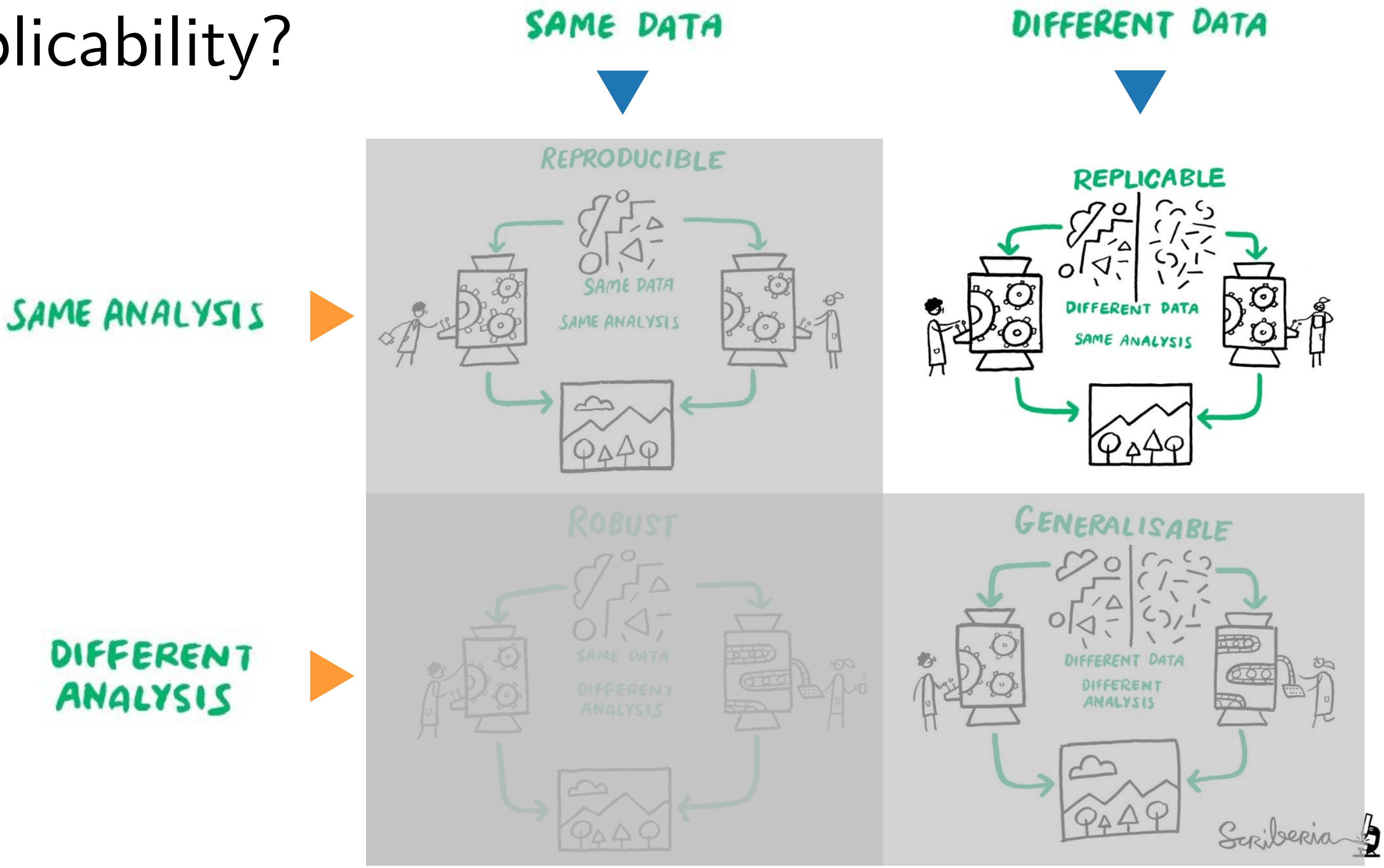
▶ TriScale

What is replicability?



The Turing Way project illustration by Scriberia. Zenodo. http://doi.org/10.5281/zenodo.3332807

What is replicability?



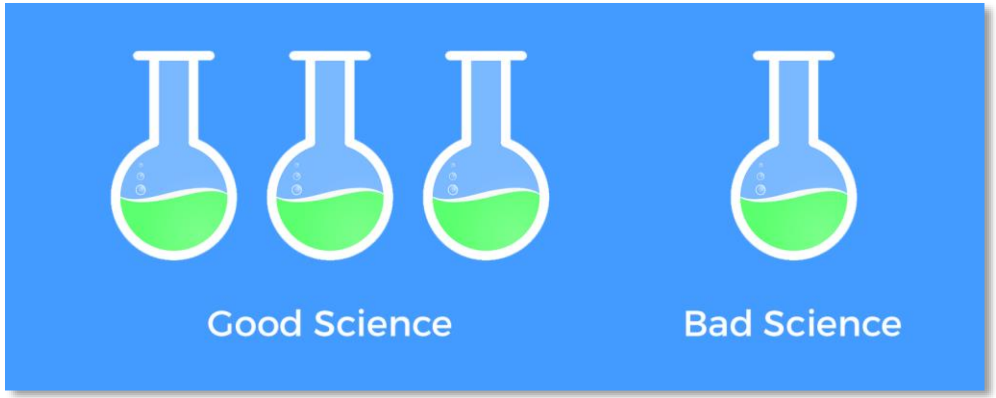
The Turing Way project illustration by Scriberia. Zenodo. http://doi.org/10.5281/zenodo.3332807

What is replicability?
Why does it matter?

Because

No result is “science” if it cannot be independently replicated by others.

In picture



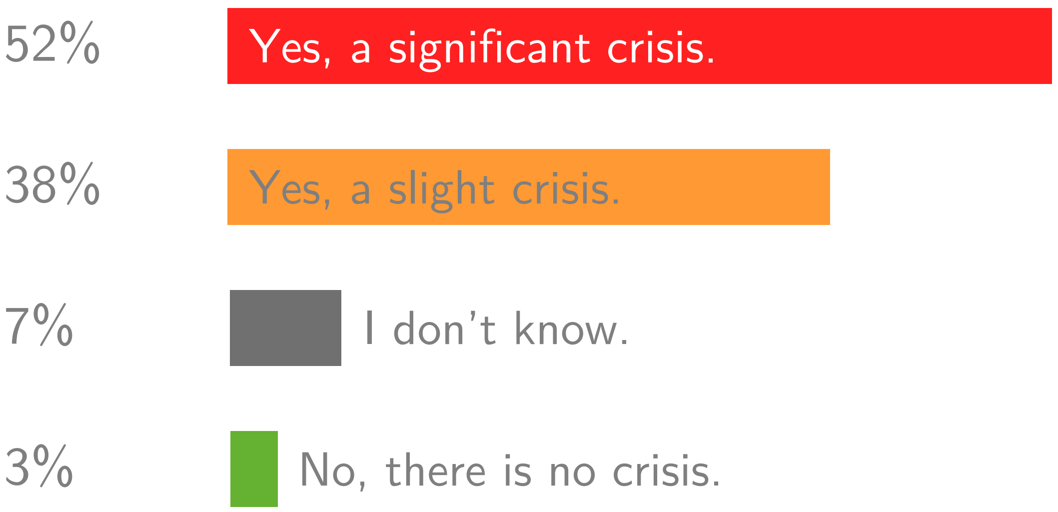
www.zbw-mediatalk.eu

“Is there a reproducibility crisis?”

Poor/no documentation
Artifacts not available
Unstable environment
Analytical bias
Falsification
etc.

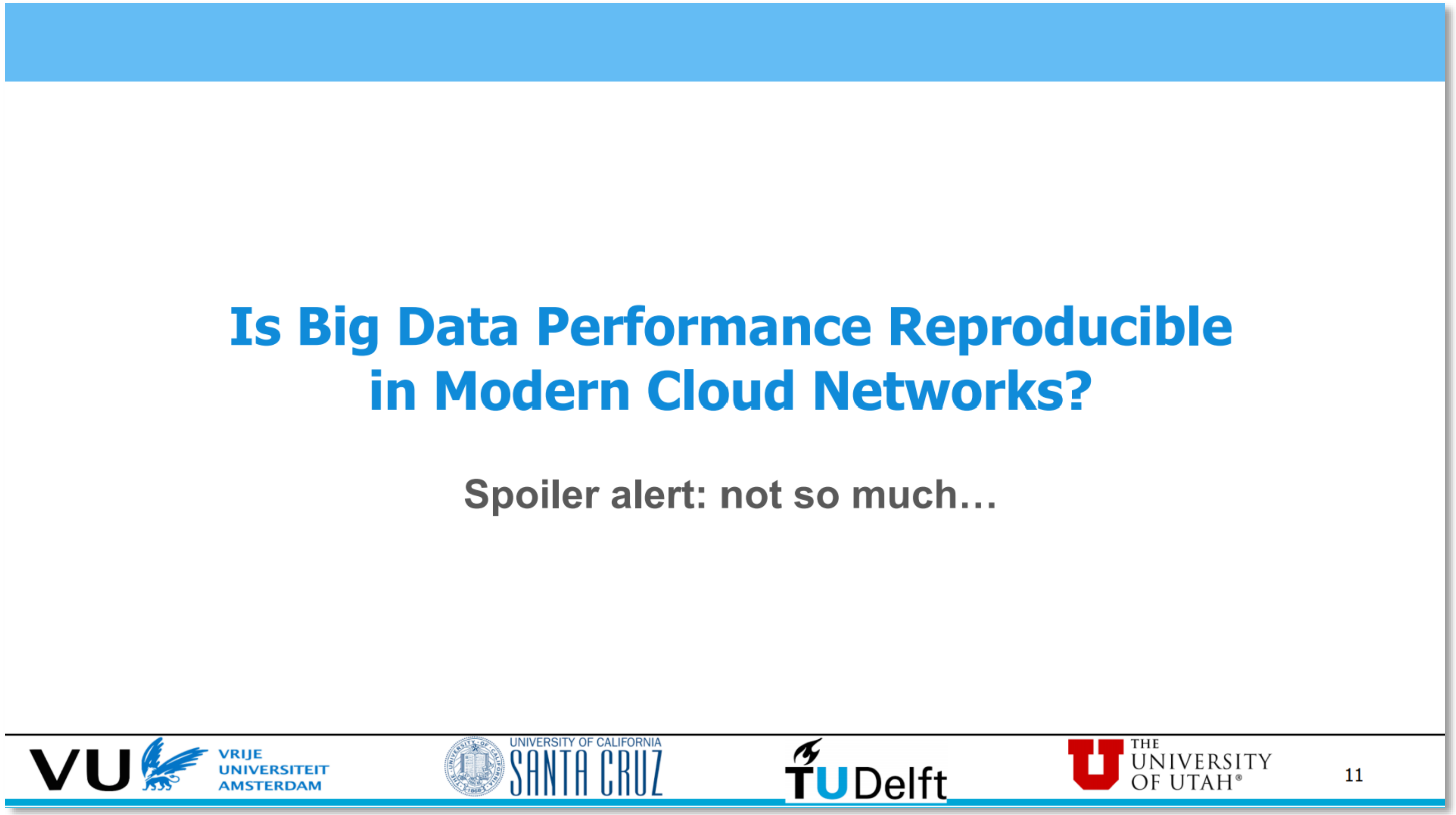
90%

of surveyed scientists stated that there is a reproducibility crisis in their research field.



Is There a Reproducibility Crisis?
Monya Baker. Nature News (2016)

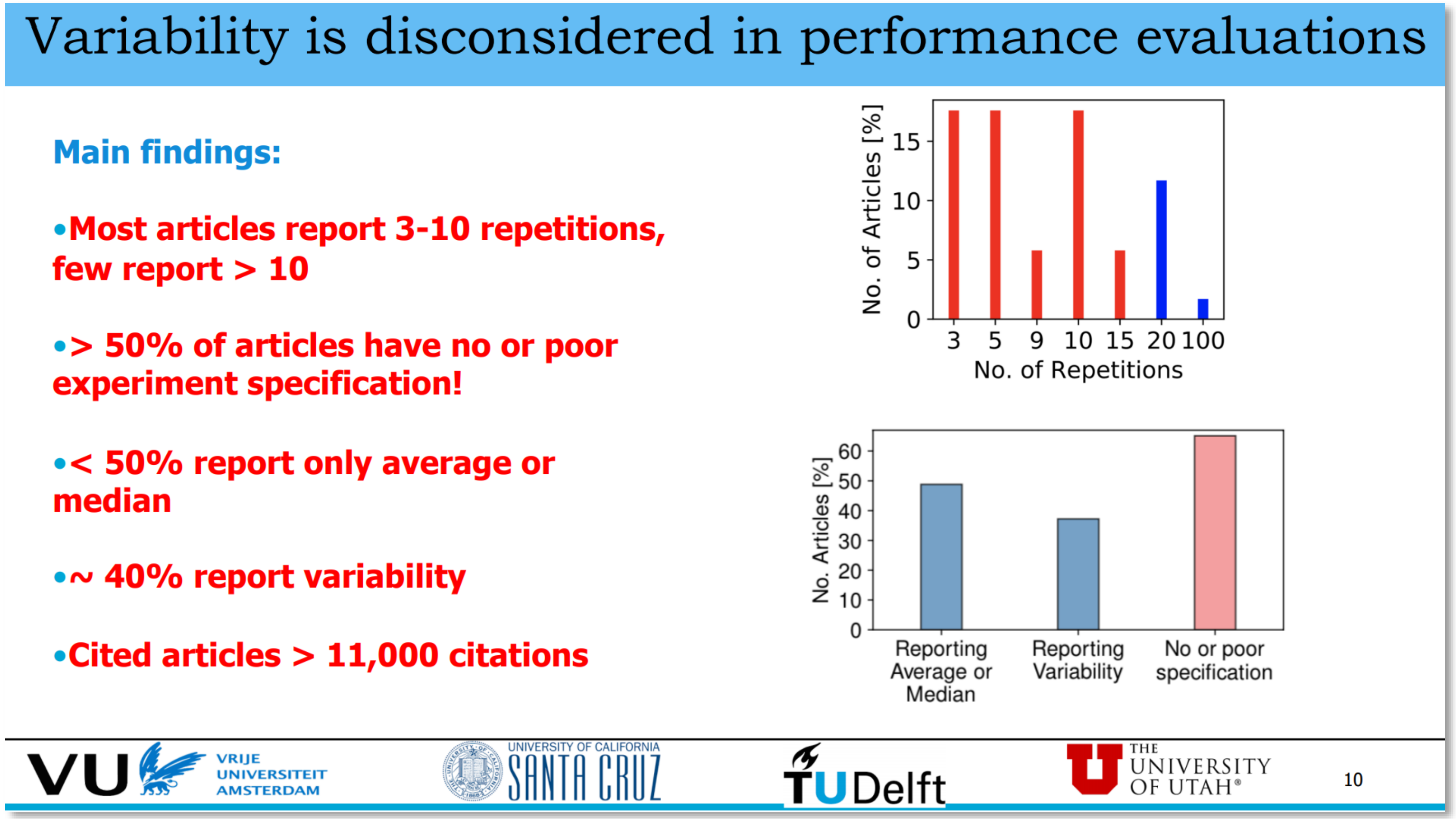
“Is there a reproducibility crisis?”
Does it really affect CS? Networking?



Uta et al., NSDI 2020

“Is there a reproducibility crisis?”

Does it really affect CS? Networking?

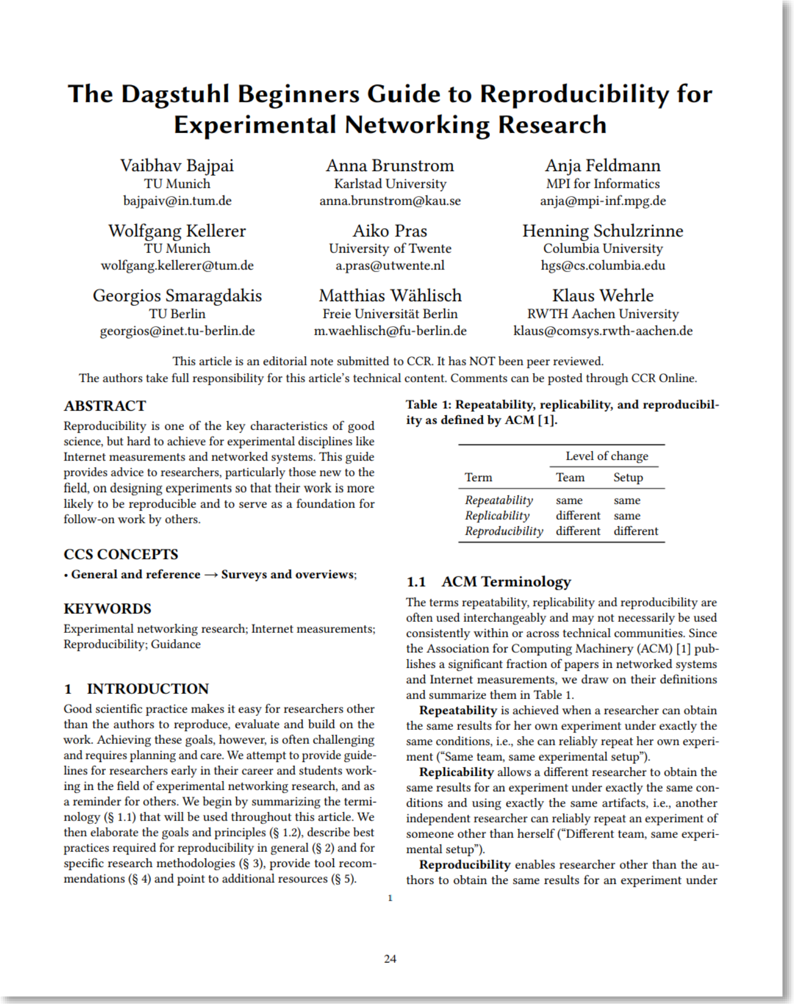


Uta et al., NSDI 2020

The literature addresses replicability issues

Two examples

- Mainly guidelines



The Dagstuhl Beginners Guide to Reproducibility for Experimental Networking Research

Vaibhav Bajpai, TU Munich, bajpaiv@in.tum.de
Anna Brunstrom, Karlstad University, anna.brunstrom@kau.se
Anja Feldmann, MPI for Informatics, anja@mpi-inf.mpg.de
Wolfgang Kellerer, TU Munich, wolfgang.kellerer@tum.de
Aiko Pras, University of Twente, a.pras@utwente.nl
Henning Schulzrinne, Columbia University, hgs@cs.columbia.edu
Georgios Smaragdakis, TU Berlin, georgios@inet.tu-berlin.de
Matthias Wählisch, Freie Universität Berlin, m.waehlisch@fu-berlin.de
Klaus Wehrle, RWTH Aachen University, klaus@comsys.rwth-aachen.de

This article is an editorial note submitted to CCR. It has NOT been peer reviewed.
The authors take full responsibility for this article's technical content. Comments can be posted through CCR Online.

ABSTRACT

Reproducibility is one of the key characteristics of good science, but hard to achieve for experimental disciplines like Internet measurements and networked systems. This guide provides advice to researchers, particularly those new to the field, on designing experiments so that their work is more likely to be reproducible and to serve as a foundation for follow-on work by others.

CCS CONCEPTS

• General and reference → Surveys and overviews;

KEYWORDS

Experimental networking research; Internet measurements; Reproducibility; Guidance

1 INTRODUCTION

Good scientific practice makes it easy for researchers other than the authors to reproduce, evaluate and build on the work. Achieving these goals, however, is often challenging and requires planning and care. We attempt to provide guidelines for researchers early in their career and students working in the field of experimental networking research, and as a reminder for others. We begin by summarizing the terminology (§ 1.1) that will be used throughout this article. We then elaborate the goals and principles (§ 1.2), describe best practices required for reproducibility in general (§ 2) and for specific research methodologies (§ 3), provide tool recommendations (§ 4) and point to additional resources (§ 5).

Table 1: Repeatability, replicability, and reproducibility as defined by ACM [1].

	Level of change	
Term	Team	Setup
Repeatability	same	same
Replicability	different	same
Reproducibility	different	different

1.1 ACM Terminology

The terms repeatability, replicability and reproducibility are often used interchangeably and may not necessarily be used consistently within or across technical communities. Since the Association for Computing Machinery (ACM) [1] publishes a significant fraction of papers in networked systems and Internet measurements, we draw on their definitions and summarize them in Table 1.

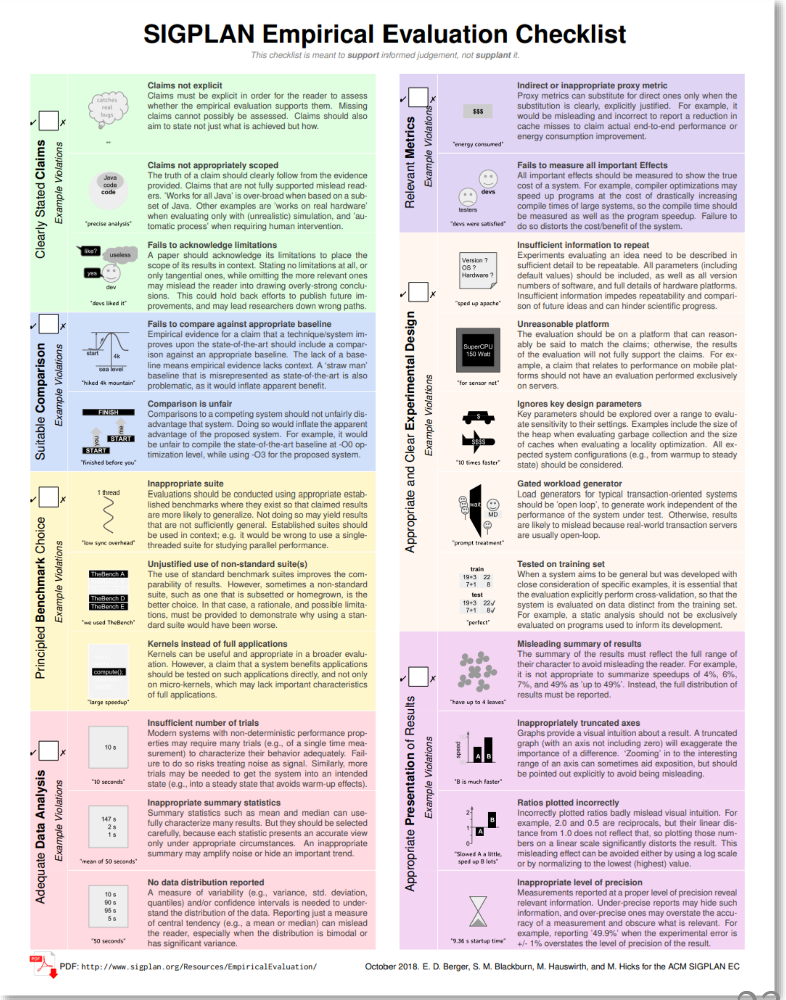
Repeatability is achieved when a researcher can obtain the same results for her own experiment under exactly the same conditions, i.e., she can reliably repeat her own experiment ("Same team, same experimental setup").

Replicability allows a different researcher to obtain the same results for an experiment under exactly the same conditions and using exactly the same artifacts, i.e., another independent researcher can reliably repeat an experiment of someone other than herself ("Different team, same experimental setup").

Reproducibility enables researcher other than the authors to obtain the same results for an experiment under

1

24



SIGPLAN Empirical Evaluation Checklist

This checklist is meant to **support** informed judgement, not **supplant** it.

Clearly Stated Claims — Example Violations
- Claims not explicit
- Claims not appropriately scoped
- Fails to acknowledge limitations

Suitable Comparison — Example Violations
- Fails to compare against appropriate baseline
- Comparison is unfair

Principled Benchmark Choice — Example Violations
- Inappropriate suite
- Unjustified use of non-standard suite(s)
- Kernels instead of full applications

Adequate Data Analysis — Example Violations
- Insufficient number of trials
- Inappropriate summary statistics
- No data distribution reported

Relevant Metrics — Example Violations
- Indirect or inappropriate proxy metric
- Fails to measure all important Effects

Appropriate and Clear Experimental Design — Example Violations
- Insufficient information to repeat
- Unreasonable platform
- Ignores key design parameters
- Gated workload generator
- Tested on training set

Appropriate Presentation of Results — Example Violations
- Misleading summary of results
- Inappropriately truncated axes
- Ratios plotted incorrectly
- Inappropriate level of precision

PDF: http://www.sigplan.org/Resources/EmpiricalEvaluation/

October 2018. E. D. Berger, S. M. Blackburn, M. Hauswirth, and M. Hicks for the ACM SIGPLAN EC

The literature addresses replicability issues but it lacks concrete answers to practical questions

For example

- How many times should one repeat an experiment?
- Which statistical methods should one use to synthesize results?

In other words

We lack a concrete methodology for the design and analysis of experiments.

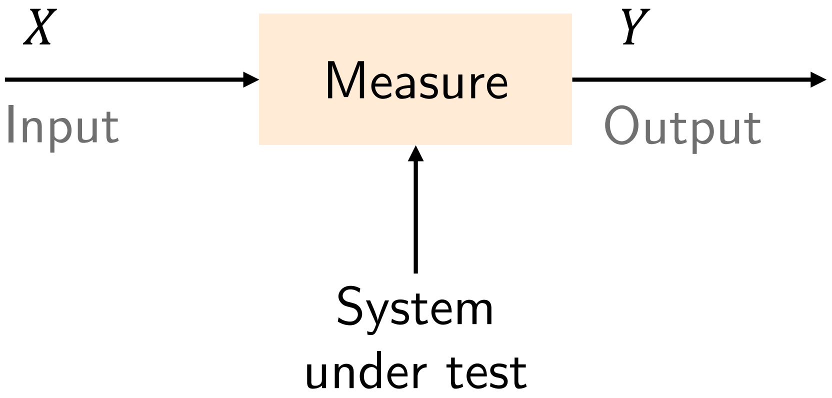
▶ That's TriScale.

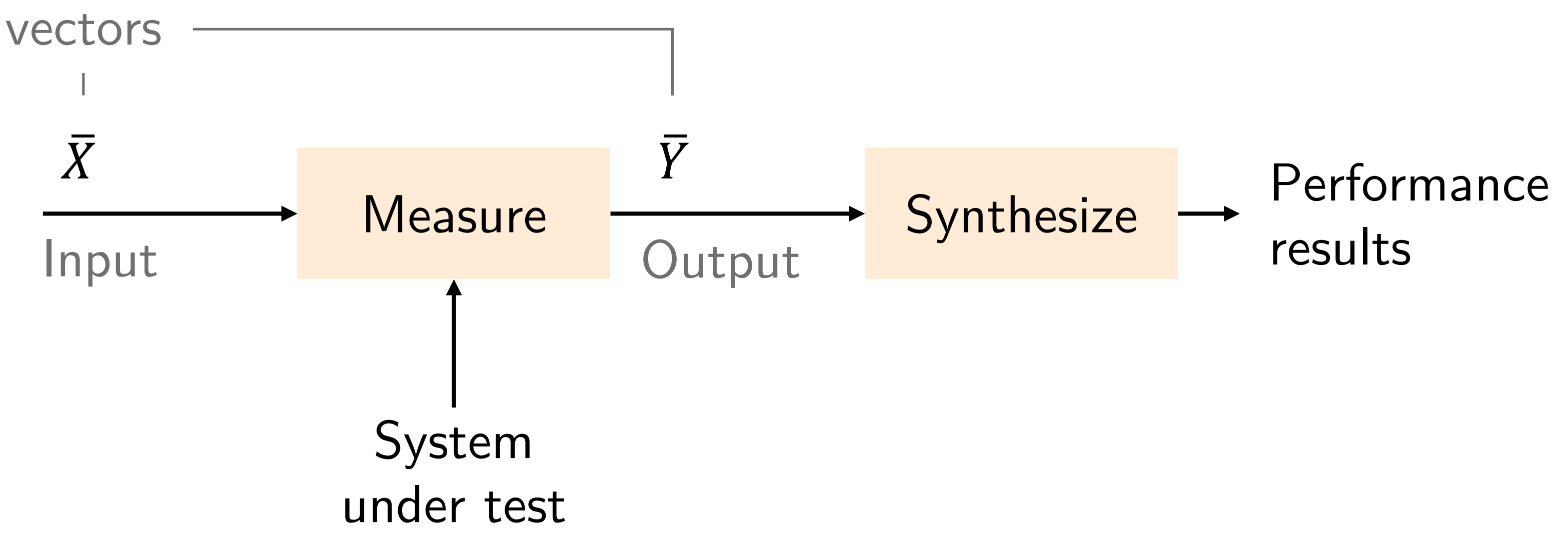
Why replicability matters
Case by example

Understanding variability
The three timescales

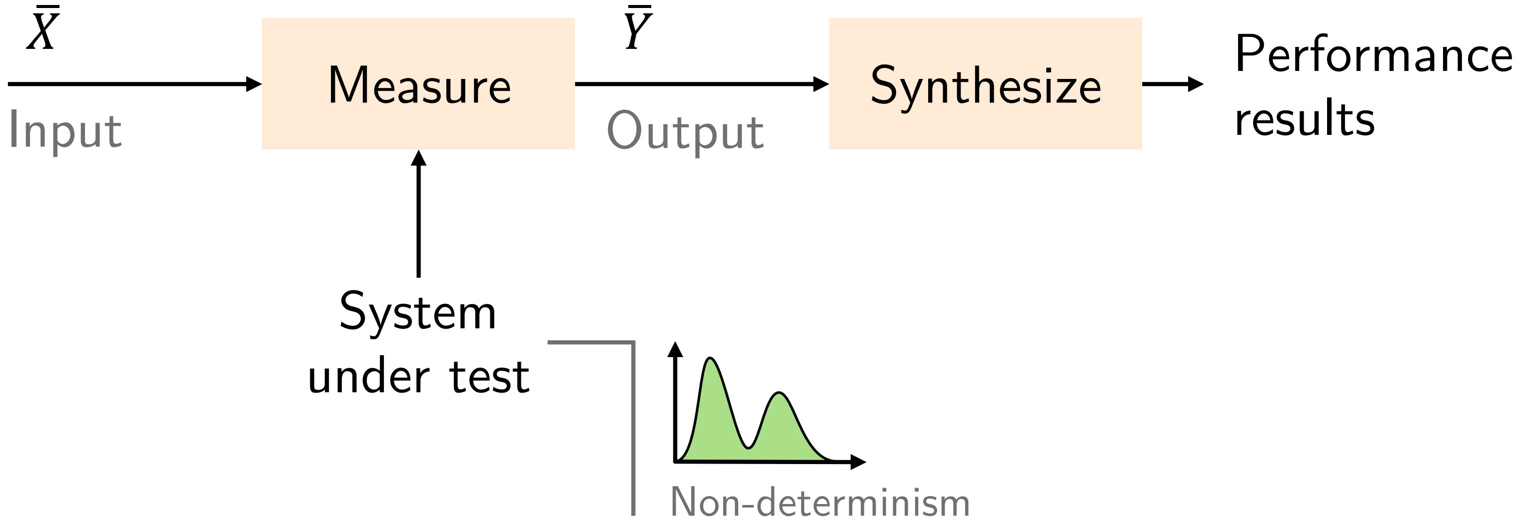
Know your data
Use the right statistics

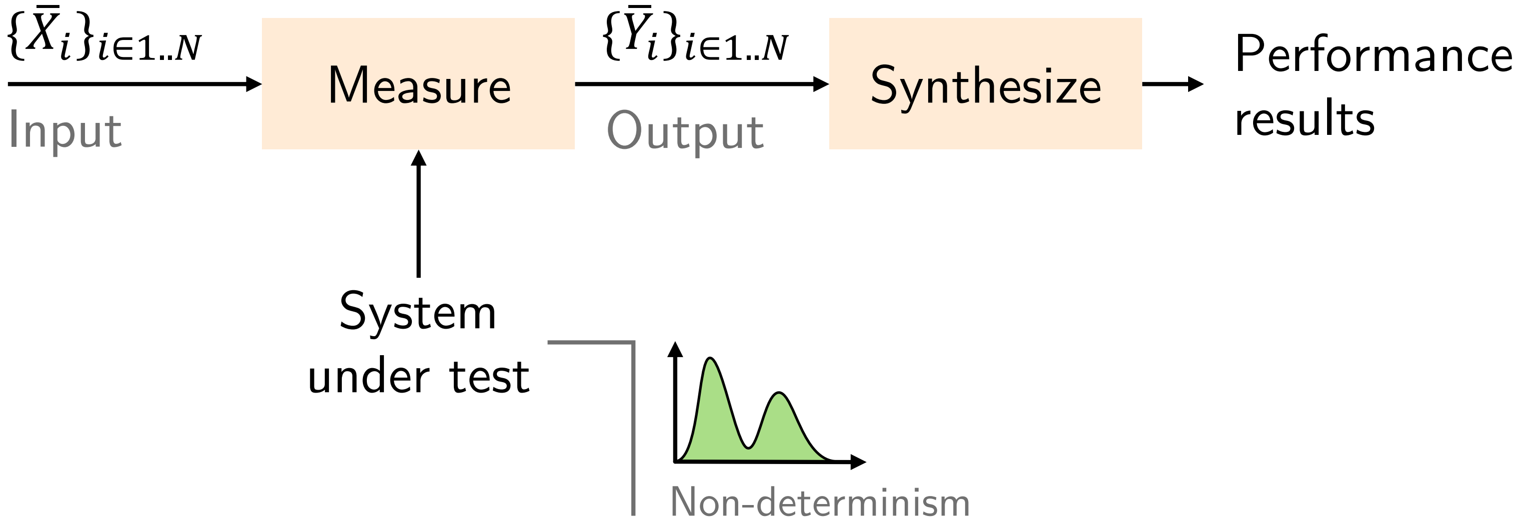
Let us have a closer look at performance evaluations

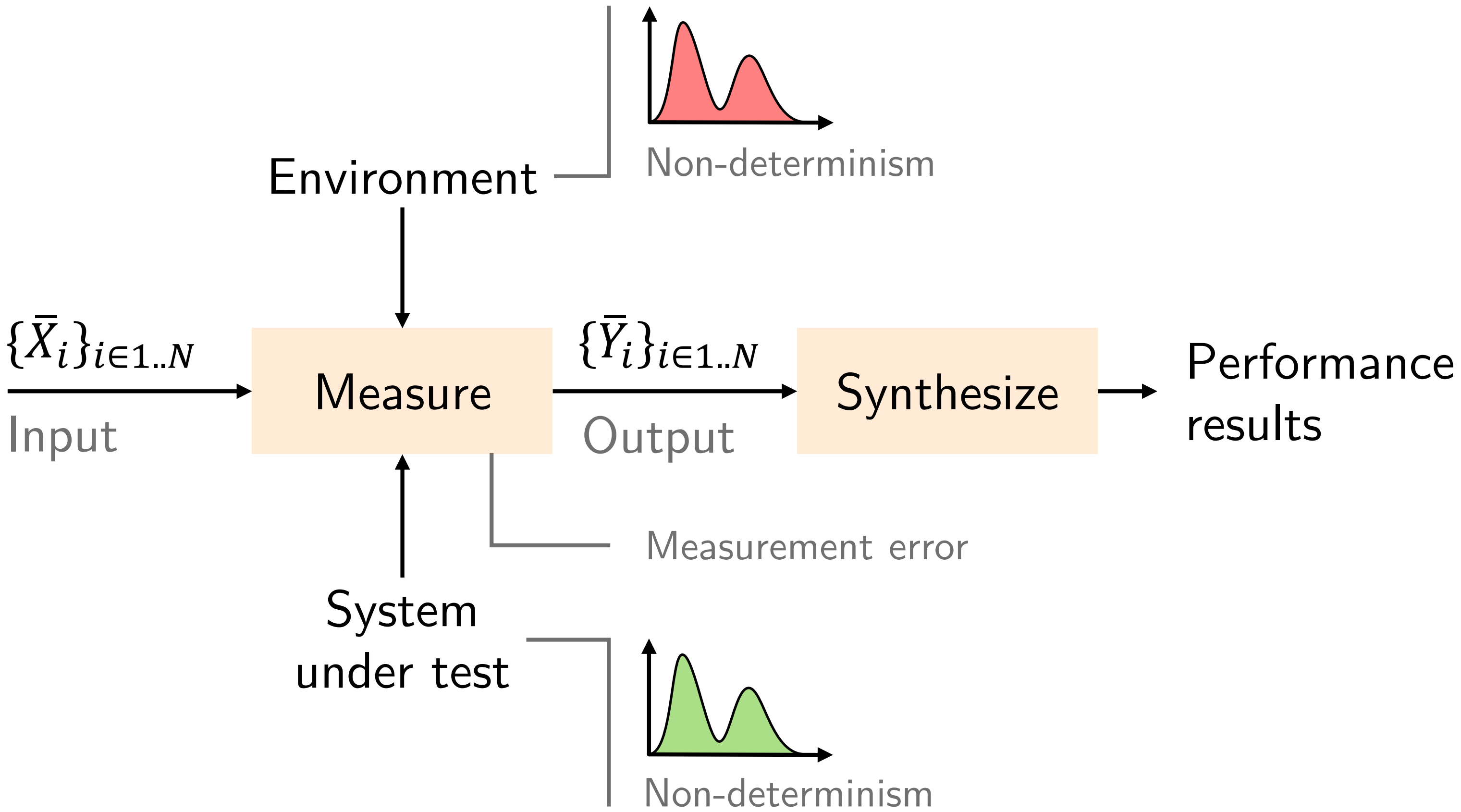

X
Y
Measure
Input
Output
System
under test

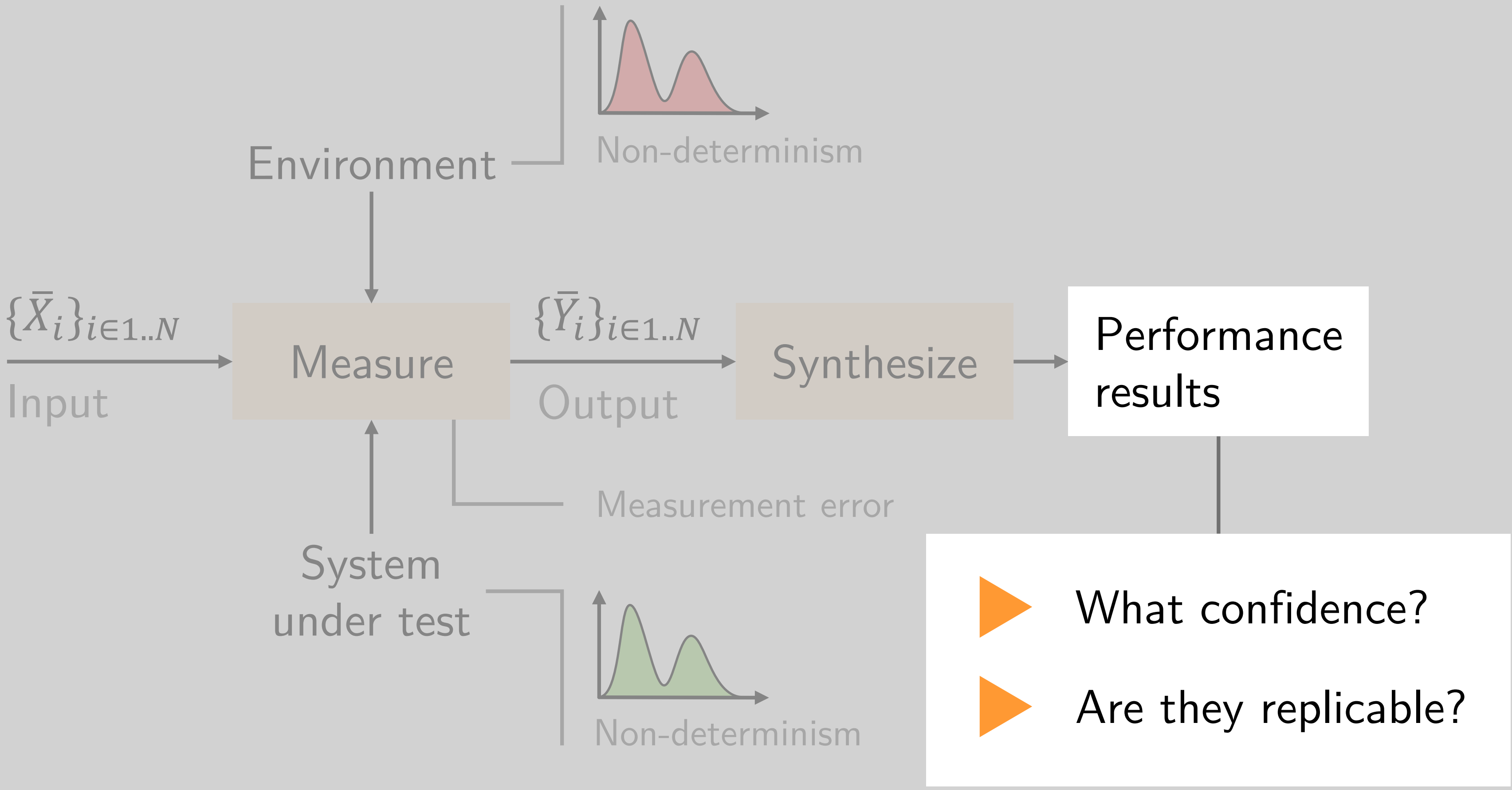


vectors
$\bar{X}$
Input
Measure
$\bar{Y}$
Output
Synthesize
Performance results
System under test

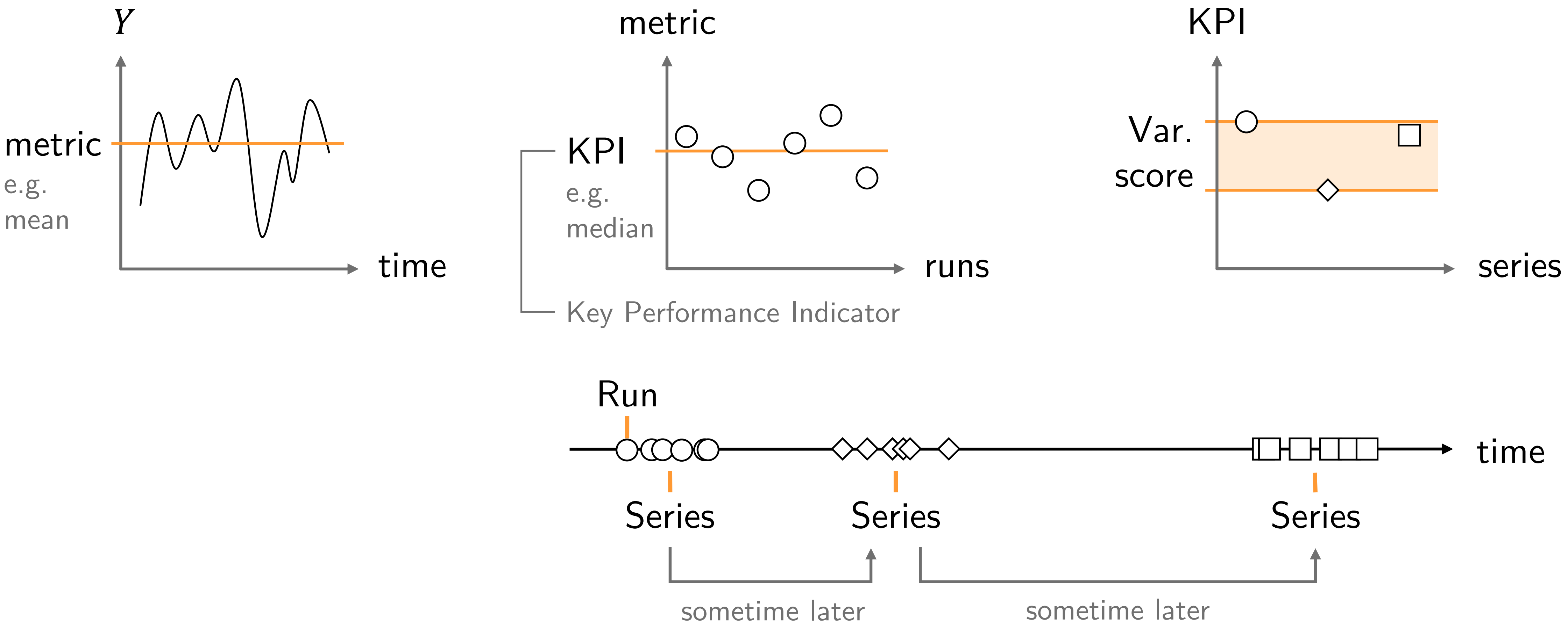

$\bar{X}$
Input
Measure
$\bar{Y}$
Output
Synthesize
Performance
results
System
under test
Non-determinism


$\{\bar{X}_i\}_{i \in 1..N}$
Input
Measure
$\{\bar{Y}_i\}_{i \in 1..N}$
Output
Synthesize
Performance results
System under test
Non-determinism

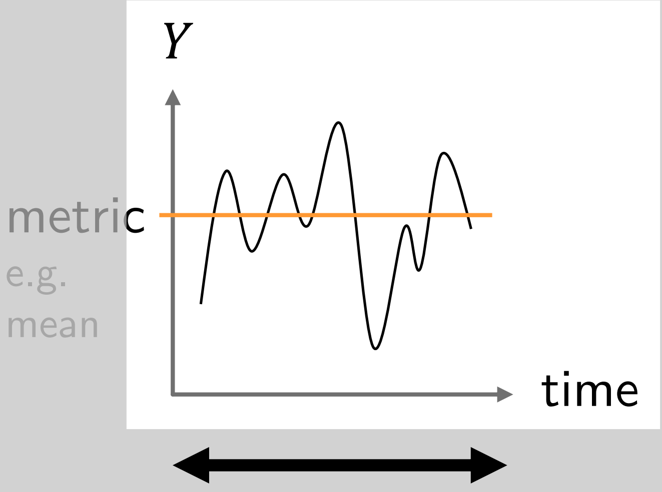

Non-determinism
Environment
$\{\bar{X}_i\}_{i \in 1..N}$
Input
Measure
$\{\bar{Y}_i\}_{i \in 1..N}$
Output
Synthesize
Performance results
Measurement error
System under test
Non-determinism


Environment
Non-determinism
$\{\bar{X}_i\}_{i \in 1..N}$
Input
Measure
$\{\bar{Y}_i\}_{i \in 1..N}$
Output
Synthesize
Performance results
Measurement error
System under test
Non-determinism
What confidence?
Are they replicable?

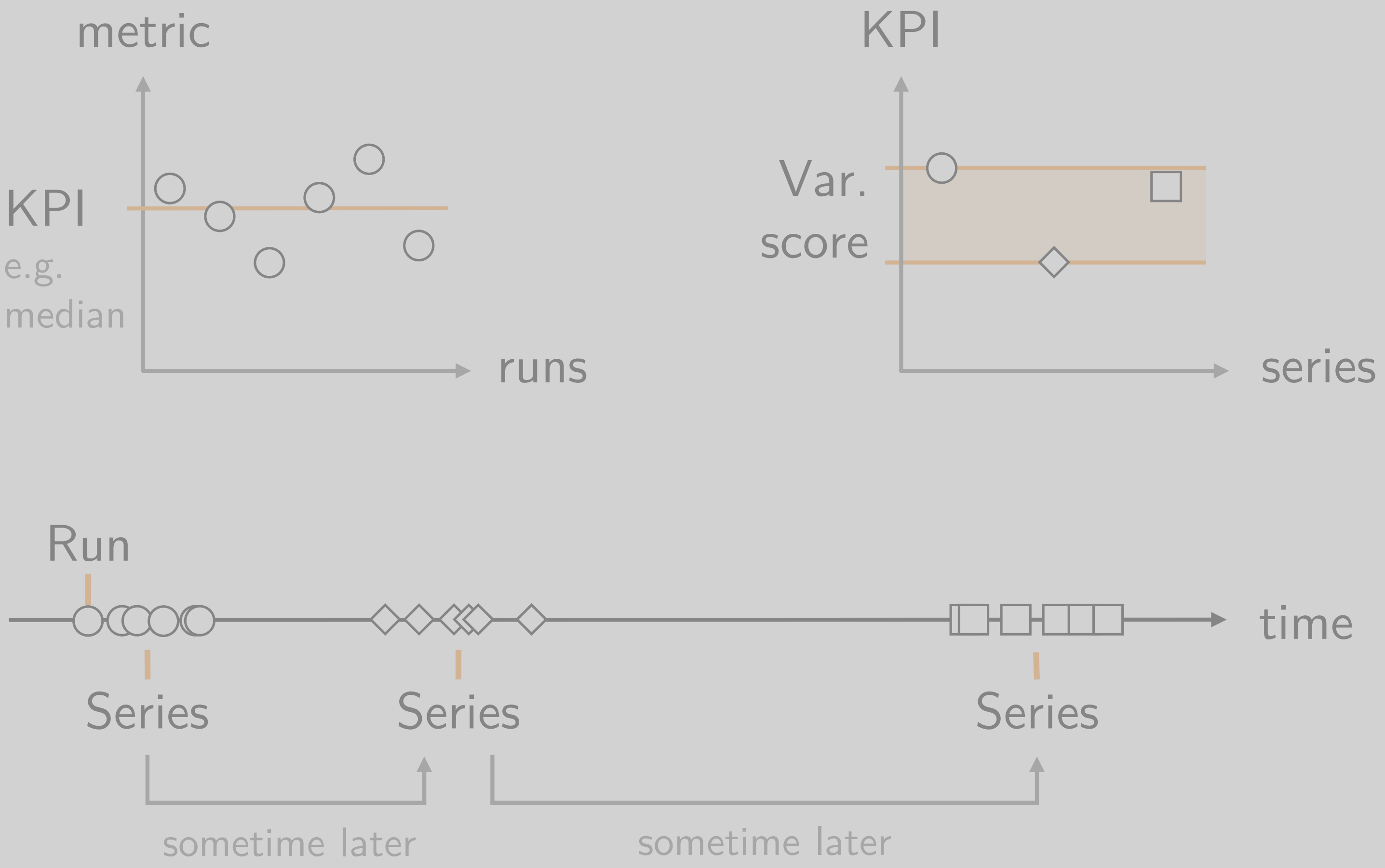
Let us set aside the causes and consider how variability looks like



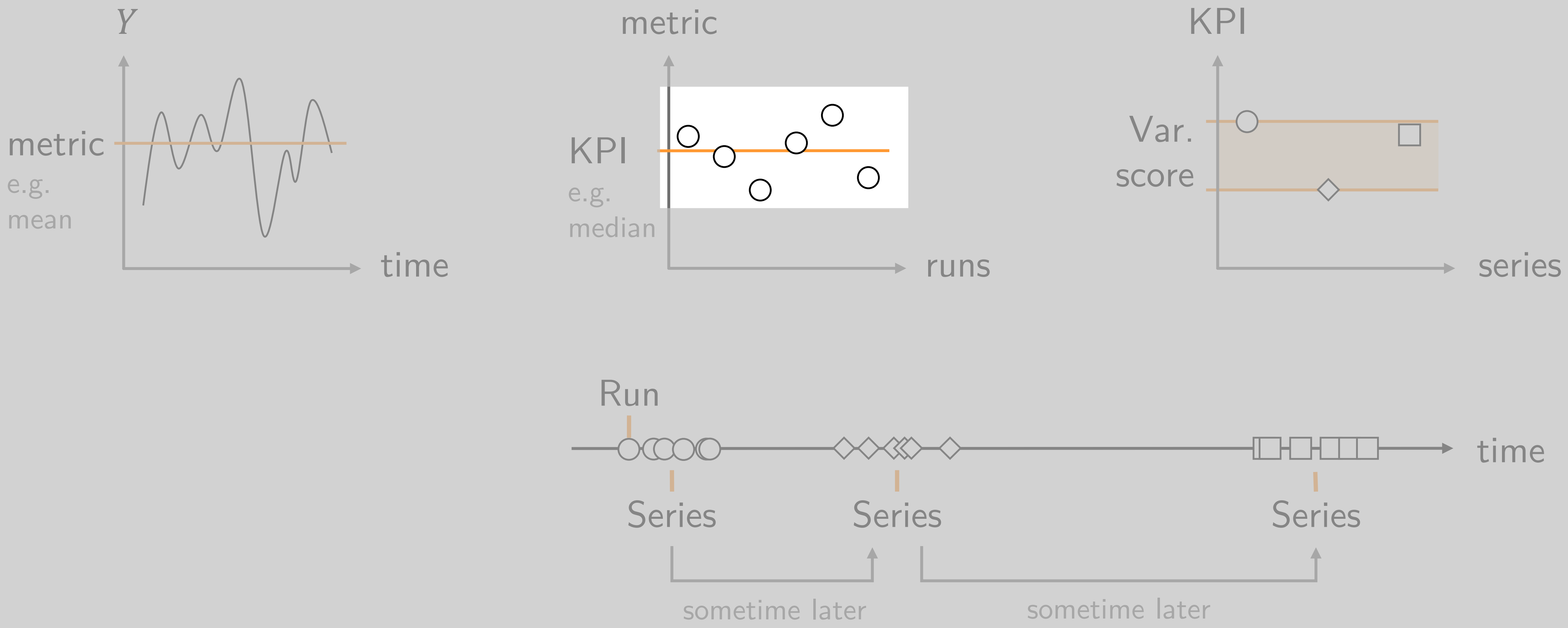
Performance varies along
three different time scales

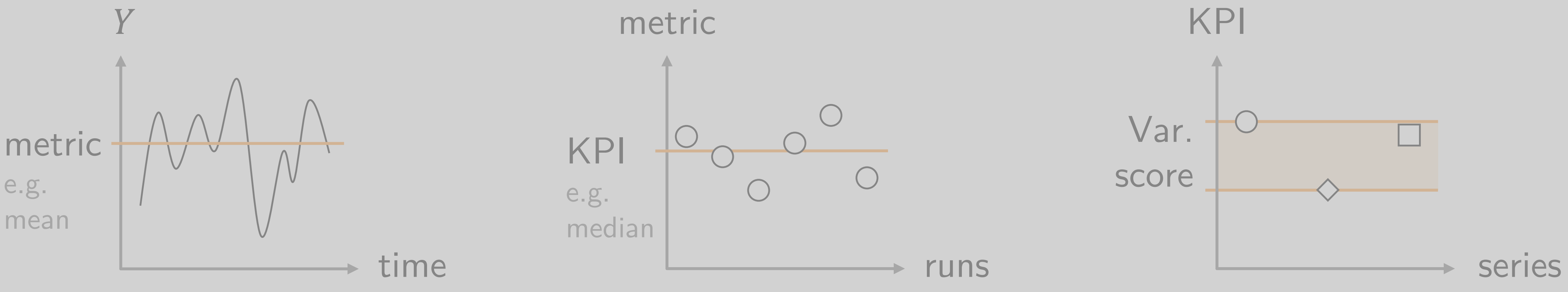


How long should a run be?



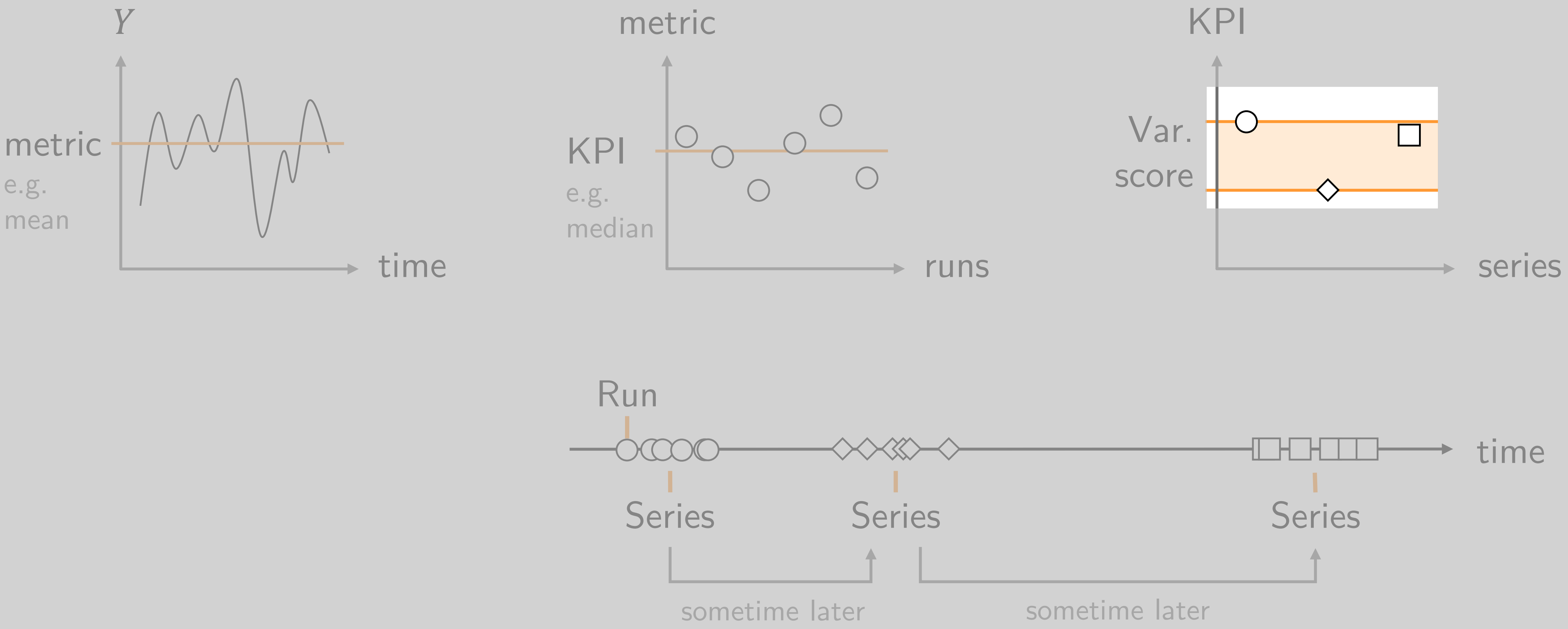
How many runs in a series?





What time span for a series?

How many series?



The four questions of experiment design

How long should a run be?

How many runs in a series?

What time span for a series?

How many series?

The four questions of experiment design

How long should a run be?

How many runs in a series?

What time span for a series?

How many series?

Objective

Find rational answers to these questions

Making statistical sense

Quantify the trade-off between

- experiment effort
- confidence in the results

Why replicability matters

Case by example

Understanding variability

The three timescales

Know your data

Use the right statistics

Let us review a few statistics basics

Statistic

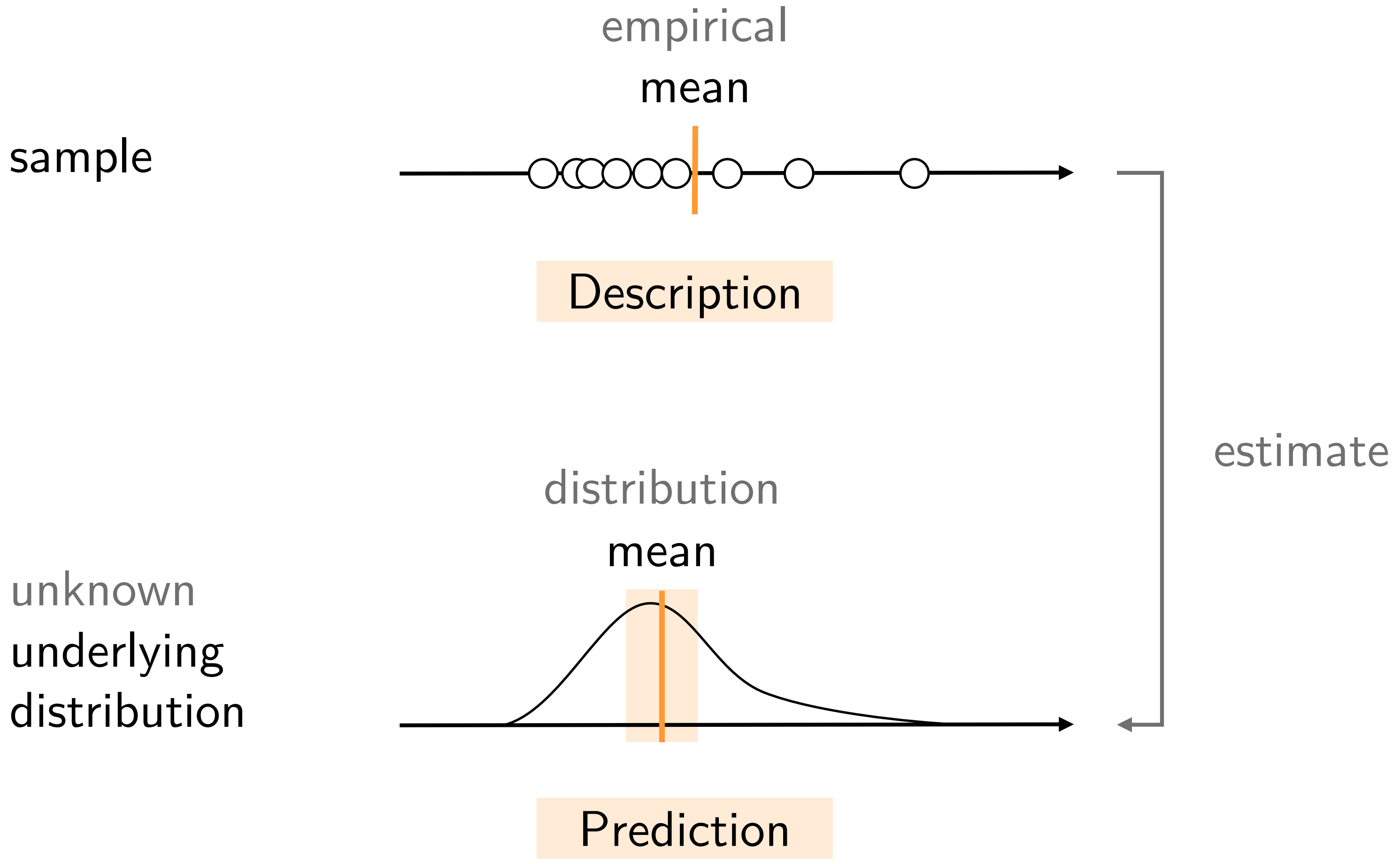
def. numerical value computed from a set of values

Let us review a few statistics basics

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

Let us review a few statistics basics

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


empirical
mean
sample
Description
estimate
distribution
mean
unknown
underlying
distribution
Prediction

Descriptive statistics $\neq$ Predictive statistics

Sample mean is X

If one draws a new sample, the sample mean is "likely" to be "close to" X

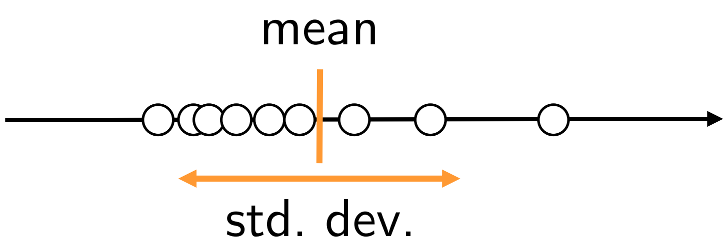
More formally?

Much stronger statement

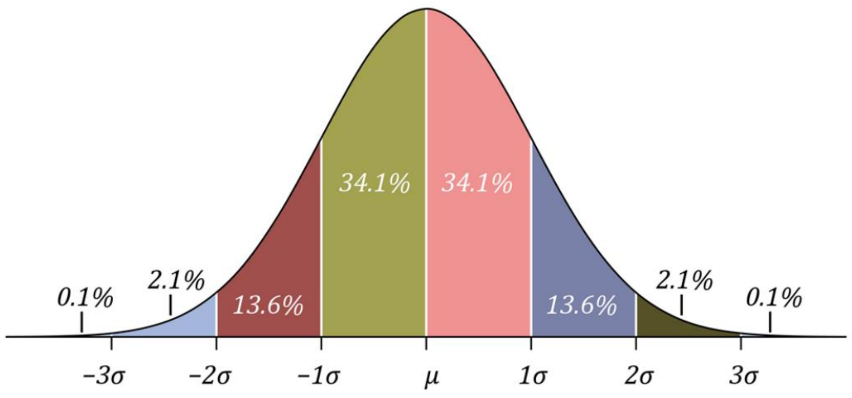
▶ Replicability!

Do these statistics say anything about the expected performance? No.

Tendency	mean
Variability	standard deviation



Prediction?



Do these statistics say anything about the expected performance? No.

If thinking so, one makes two mistakes

#1 The mean of the sample is not the mean of the underlying distribution.

Let us review a few statistics basics

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

Do these statistics say anything about the expected performance? No.

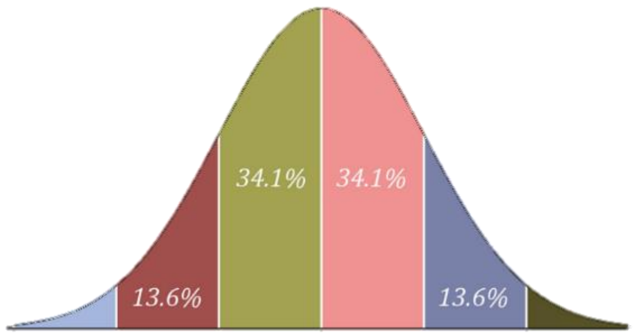
If thinking so, one makes two mistakes

#1 The mean of the sample is not the mean of the underlying distribution.

#2 The underlying distribution is not normal (almost always).

Normal

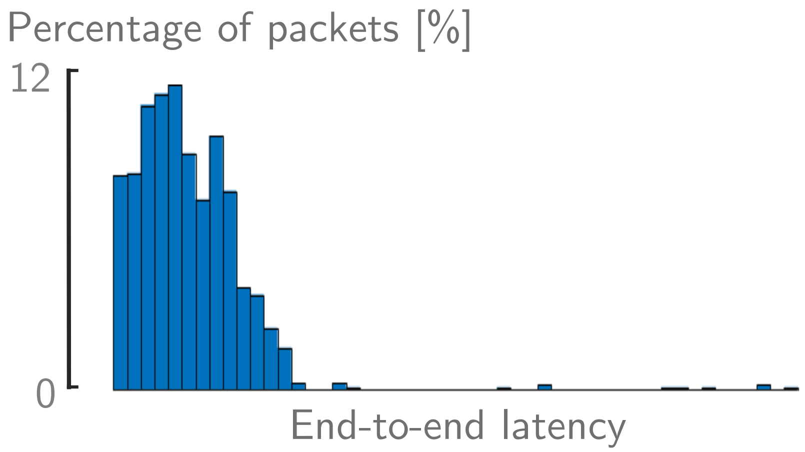
Very rare



- Cannot be assumed unless you know for sure

Not normal

Ubiquitous



Do these statistics say anything about the expetect performance? No.

If thinking so, one makes two mistakes

#1 The mean of the sample is not the mean of the underlying distribution.

- Use confidence intervals

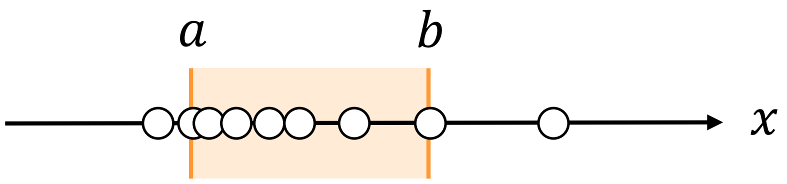
#2 The underlying distribution is not normal (almost always).

- Use non-parametric statistics

Confidence interval (CI)

Numerical interval in which lies the (unknown) true value of some parameter with a certain probability, called the confidence level

Example



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

which means that

The probability that the true median of x is within $[a, b]$ is larger or equal to 95%.

Non-parametric statistical methods

(Predictive) statistics making no assumptions on the nature of the underlying distribution

	~~Parametric~~	Non-parametric
Examples	t-test	Mann-Whitney
	ANOVA	Kruskal-Wallis
	Correlation coefficient	Spearman rank correlation
Differences	More powerful Assume normality	More general

Non-parametric statistical methods

(Predictive) statistics making no assumptions on the nature of the underlying distribution

	Parametric	Non-parametric
Examples	t-test	Mann-Whitney
	ANOVA	Kruskal-Wallis
	Correlation coefficient	Spearman rank correlation
Differences	More powerful Assume normality	More general

Statistics take-away for replicability in networking

1. Replicability requires predictive statistics

2. Predictions require confidence intervals

3. Non-parametric statistics should be used;
   do not assume normality!

Any questions?

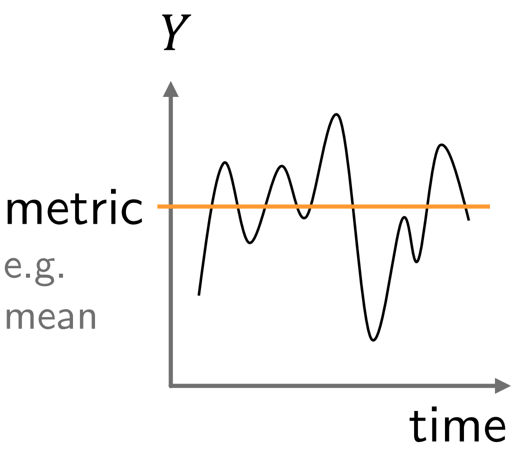
Up next

- Getting started on TriScale per se

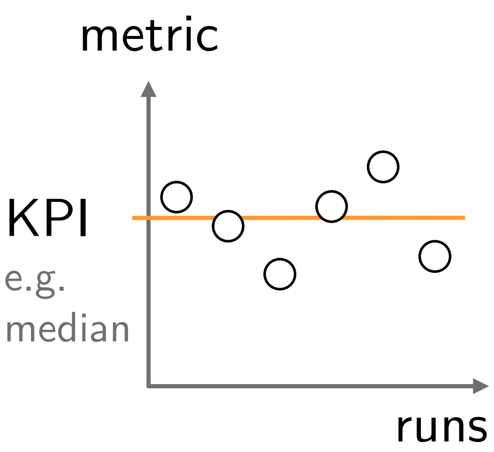
TriScale is a framework helping to design and analyze networking experiments

- Divides the experiment design and data analysis into three time scales

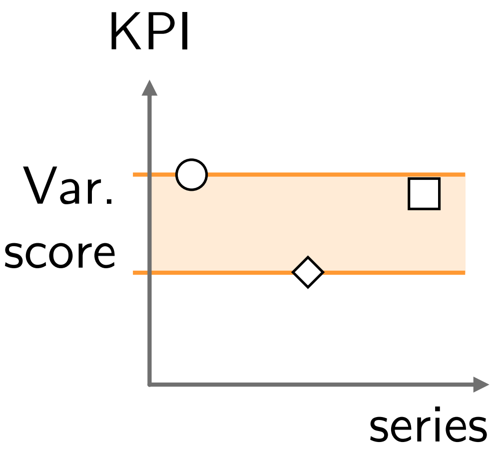
Runs

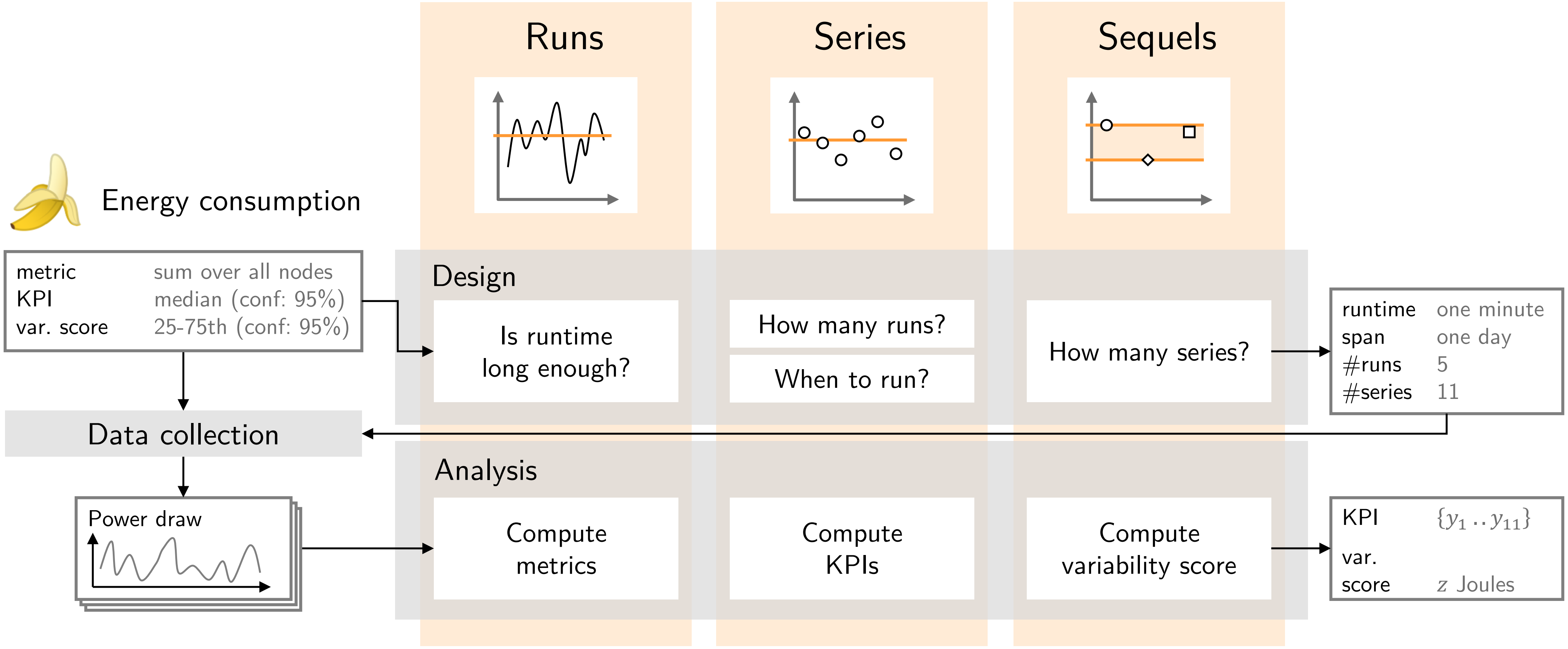


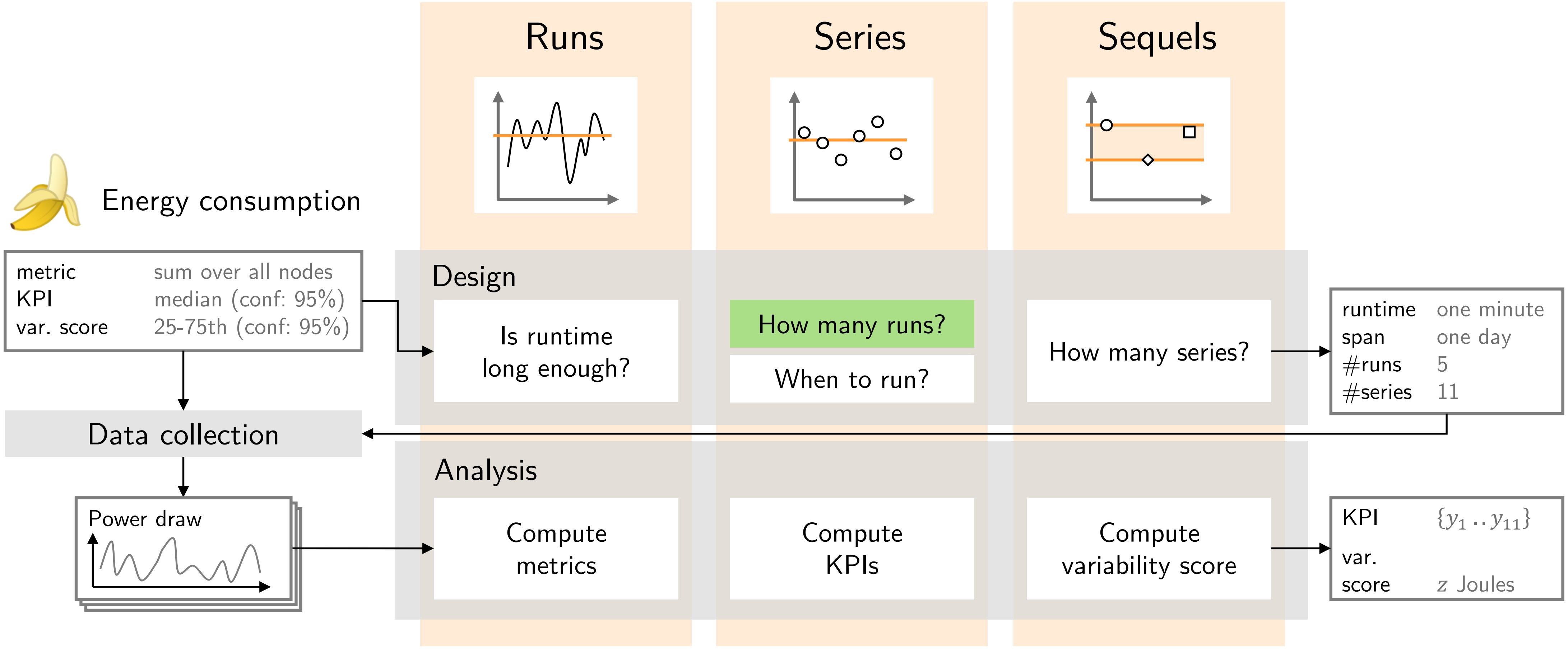
Series



Sequels

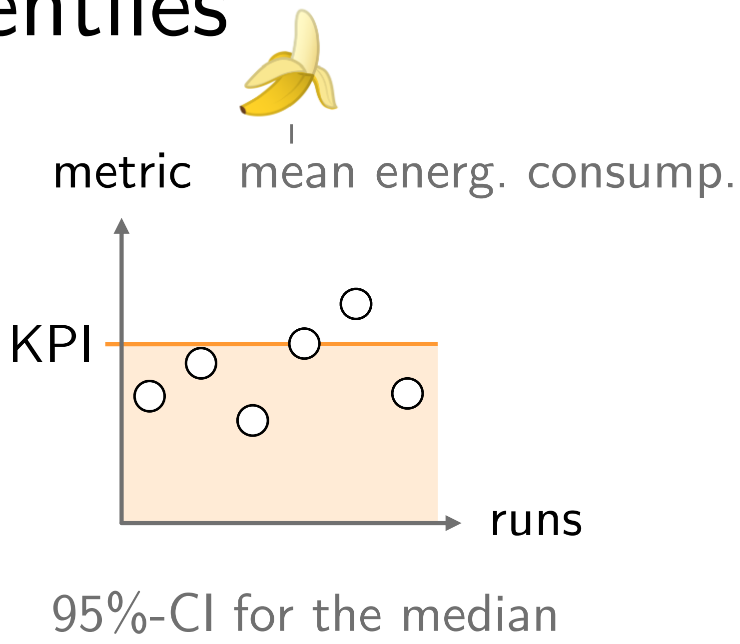



Runs
Series
Sequels
Energy consumption
metric sum over all nodes
KPI median (conf: 95%)
var. score 25-75th (conf: 95%)
Design
Is runtime long enough?
How many runs?
When to run?
How many series?
runtime one minute
span one day
#runs 5
#series 11
Data collection
Analysis
Power draw
Compute metrics
Compute KPIs
Compute variability score
KPI $\{y_1 .. y_{11}\}$
var. score z Joules



Runs
Series
Sequels
Energy consumption
metric
sum over all nodes
KPI
median (conf: 95%)
var. score
25-75th (conf: 95%)
Design
Is runtime long enough?
How many runs?
When to run?
How many series?
runtime
one minute
span
one day
#runs
5
#series
11
Data collection
Analysis
Power draw
Compute metrics
Compute KPIs
Compute variability score
KPI
$\{y_1 \mathinner{..} y_{11}\}$
var. score
z Joules

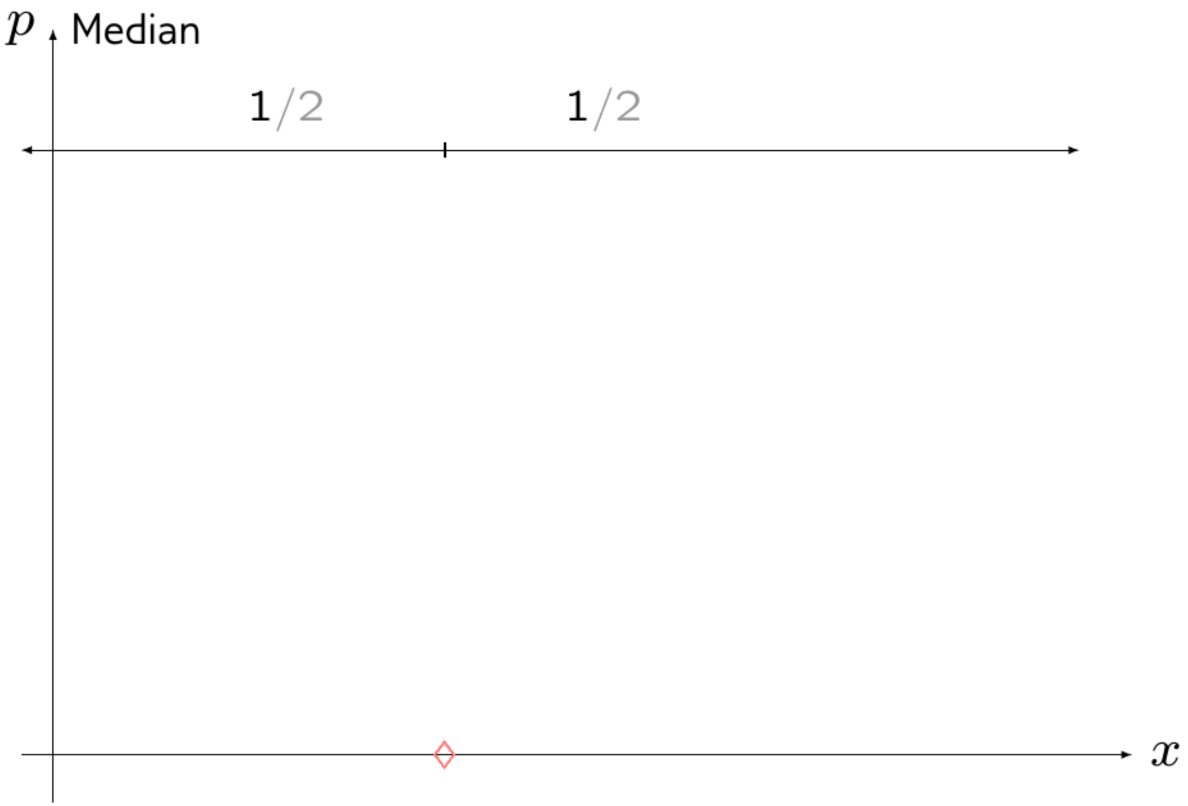
The Thompson's method provides non-parametric CI for distribution percentiles

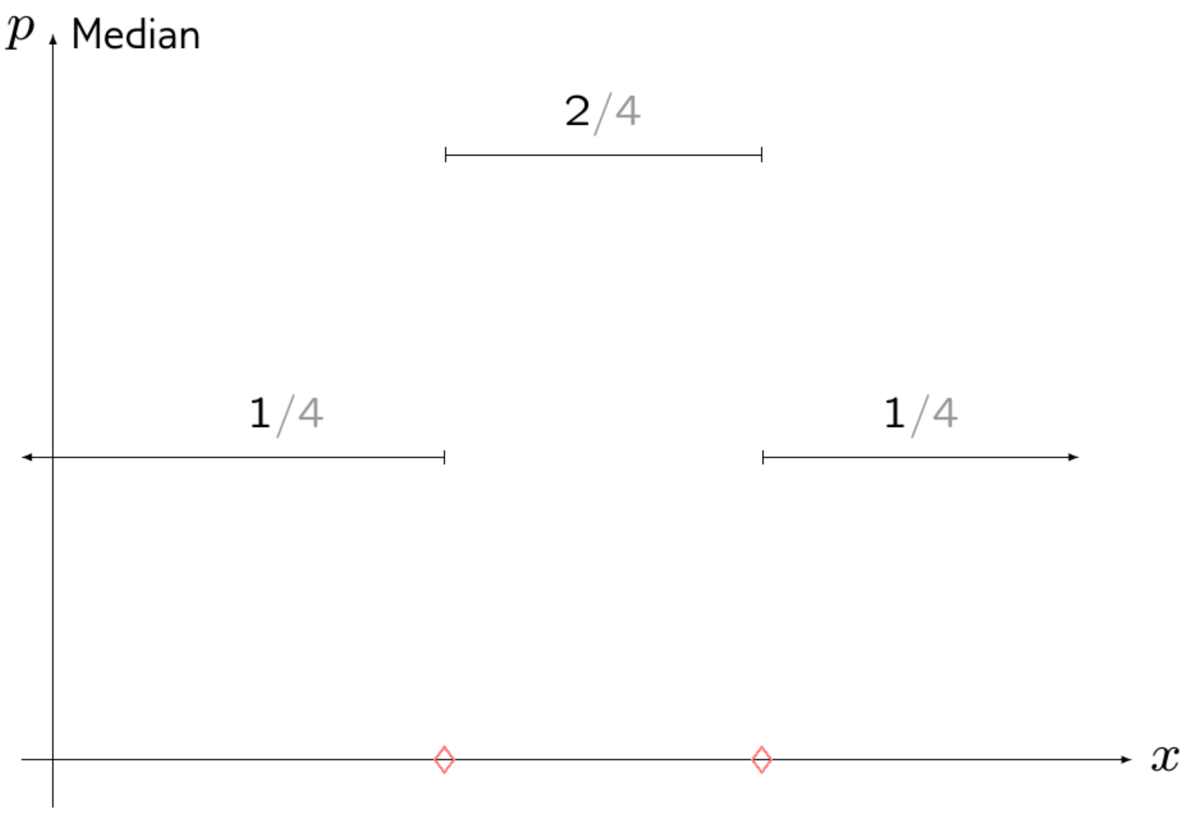


Key Performance Indicators (KPIs) are percentiles of the distribution of metric values

- Compute upper and lower bounds on the true percentile values for a certain confidence level
- The KPIs are defined as one such bounds

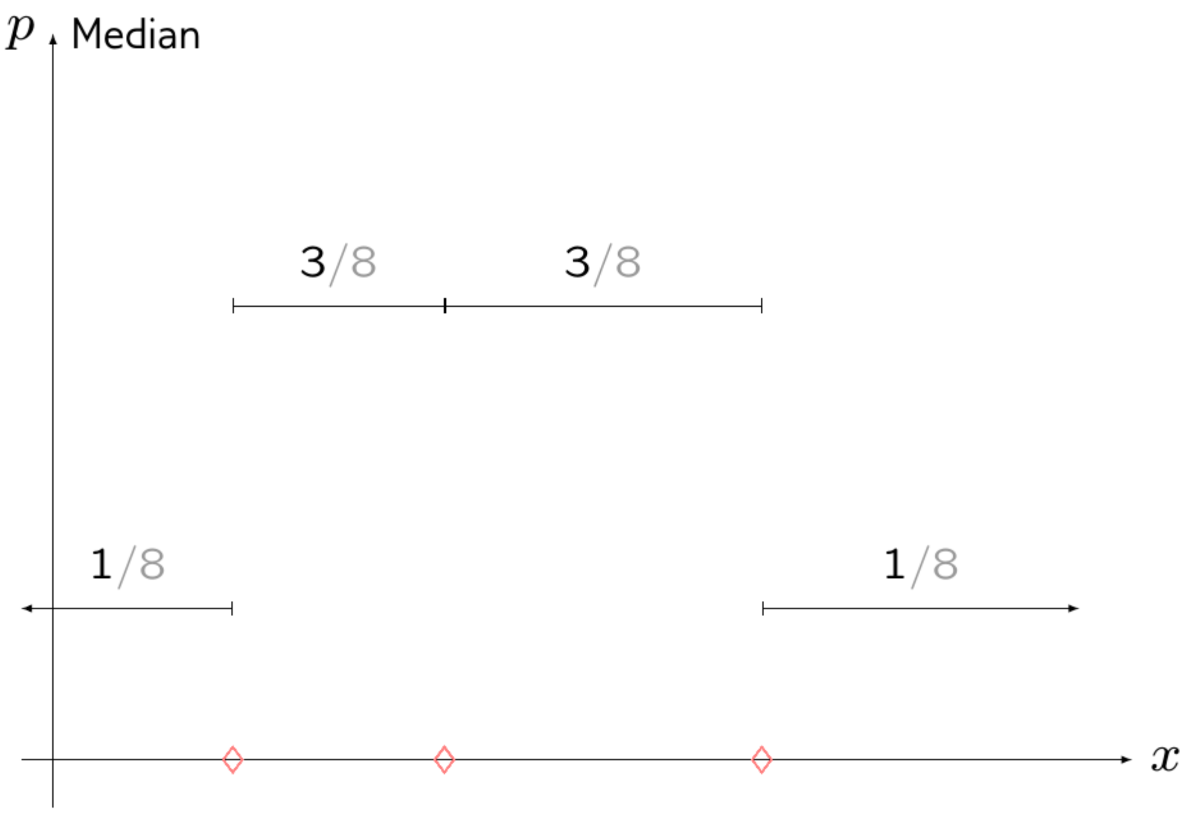
Adapted from Hanspeter Schmid and Alex Huber


p
Median
$1/2$
$1/2$
x

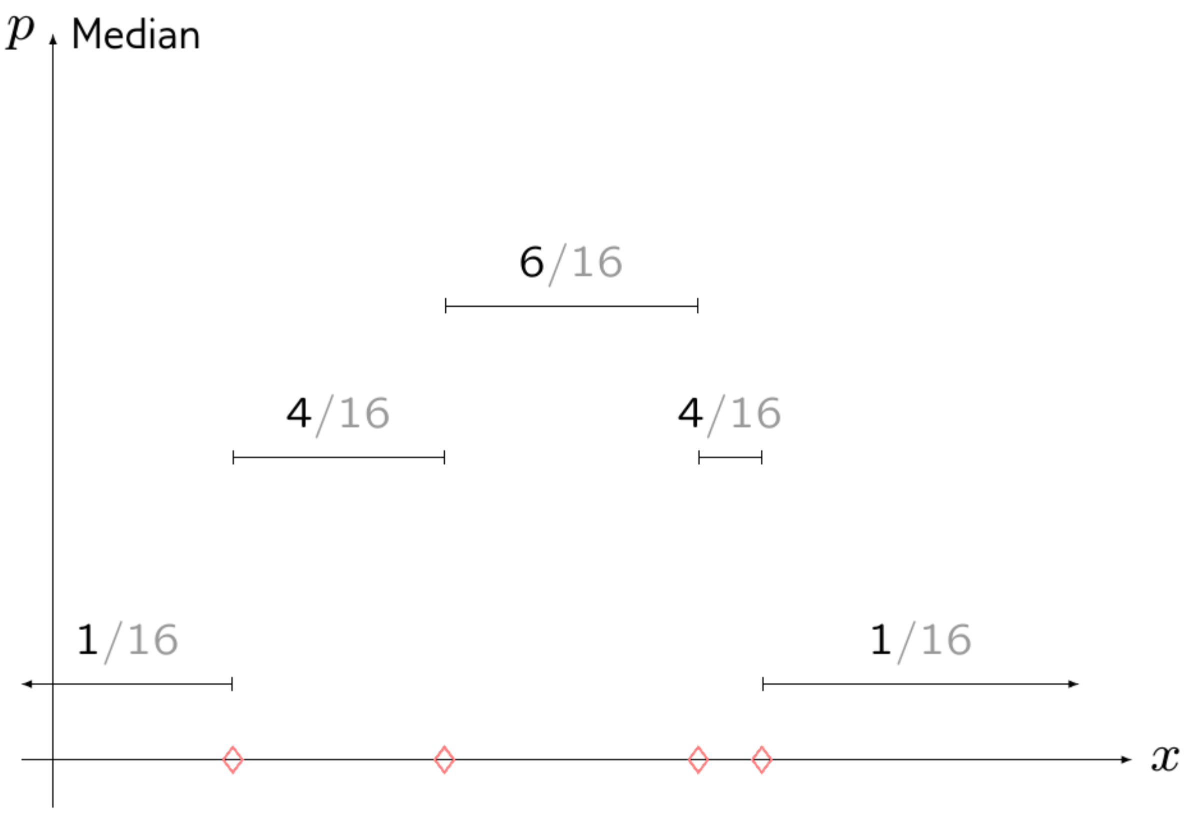


Hypothesis

Samples are i.i.d.



Hypothesis Samples are i.i.d.



Hypothesis

Samples are i.i.d.

The Thompson's method provides non-parametric CI for distribution percentiles

Probability of any P_p to be between two consecutive samples

$$\mathsf{P}\left\{x_k \le P_p \le x_{k+1}\right\} = \underbrace{\binom{N}{k} p^k (1-p)^{N-k}}_{\text{Binomial distribution}}$$

Allows to derive lower and upper bounds for any percentile

$$\mathsf{P}\left\{x_m \le P_p\right\} = 1 - \sum_{k=0}^{m-1} \binom{N}{k} p^k (1-p)^{N-k}$$

The Thompson's method provides non-parametric CI for distribution percentiles

For any confidence c

For any percentile P_p

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

For any confidence c

For any percentile P_p

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

95% CI $c = 0{,}95$	Minimal number of runs in a series
Median $p = 0{,}5$	6
25-th $p = 0{,}25$	11
1-th $p = 0{,}01$	299
0.001-th $p = 0{,}00001$	299572

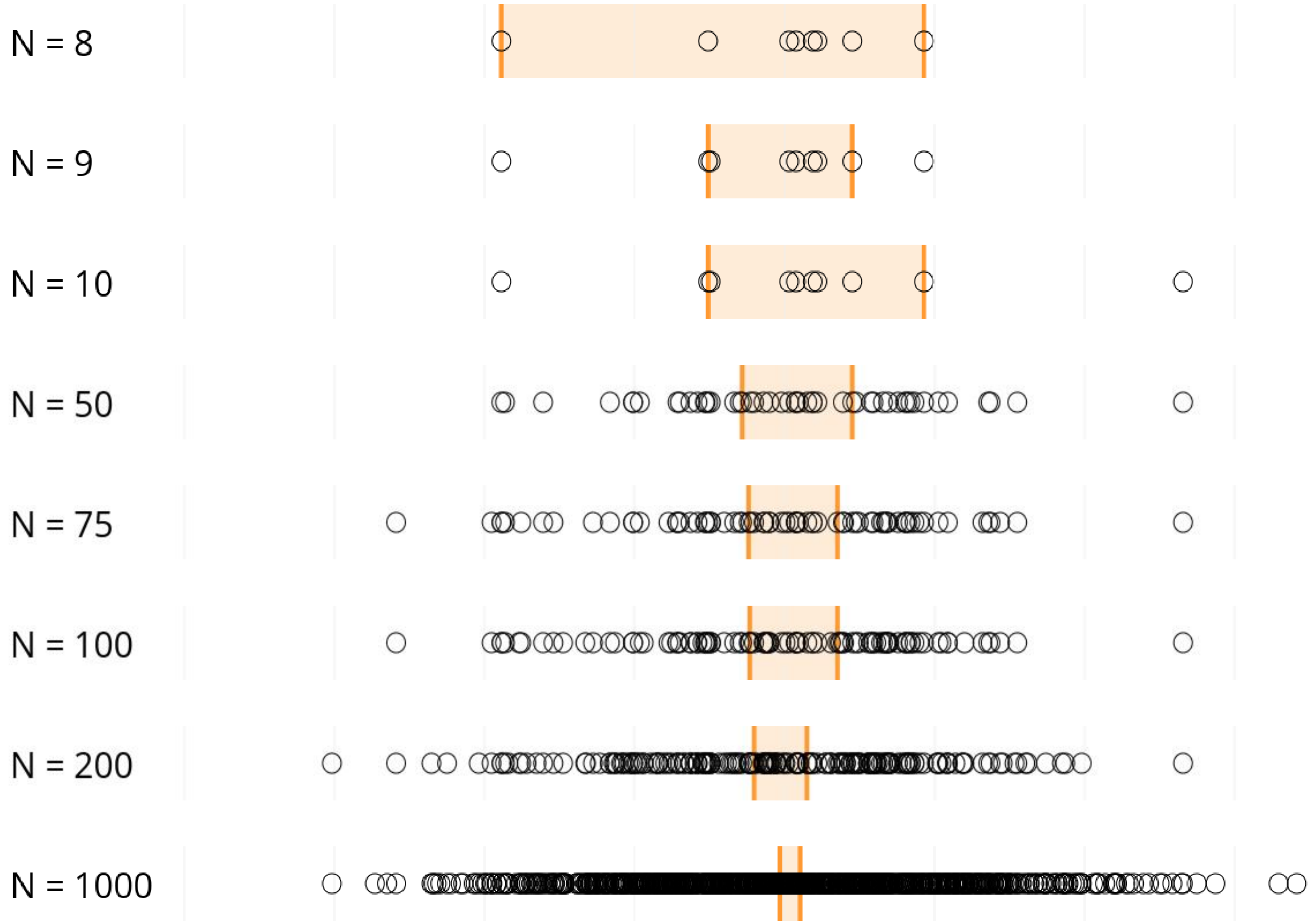
We might want to rethink
the idea of “five-nines” claims...

95%CI on the median

Minimum 6 samples

CI starts excluding most extreme values

CI gets narrower with more samples in general



Let's practice!

Go to

triscale.ethz.ch

TriScale

TriScale

A Framework Supporting Replicable Performance Evaluations in Networking

View the Project on GitHub
romain-jacob/triscale

TriScale

A Framework Supporting Replicable Performance Evaluations in Networking

Paper | Code | Tutorial | Discussion

> *Following a live tutorial session? Here are the links you're looking for*

Hands-on	
Part 1	launch binder
Part 2	launch binder

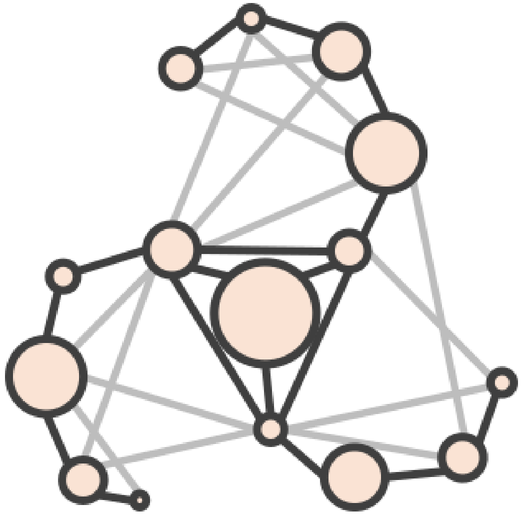
When designing their performance evaluations, networking researchers often encounter questions such as:

- How long should a run be?
- How many runs to perform?
- How to account for the variability across multiple runs?
- What statistical methods should be used to analyze the data?

Despite the best intentions, researchers often answer these questions differently, thus impairing the replicability of evaluations and the confidence in the results.

Improving the standards of replicability has recently gained traction overall, as well as within the networking community. As an important piece of the puzzle, we developed a systematic methodology that streamlines the design and analysis of performance evaluations, and we have implemented this methodology into a framework called *TriScale*.

Experimental Reproducibility in Networking Research



Resuming at 15:05
Strasbourg time

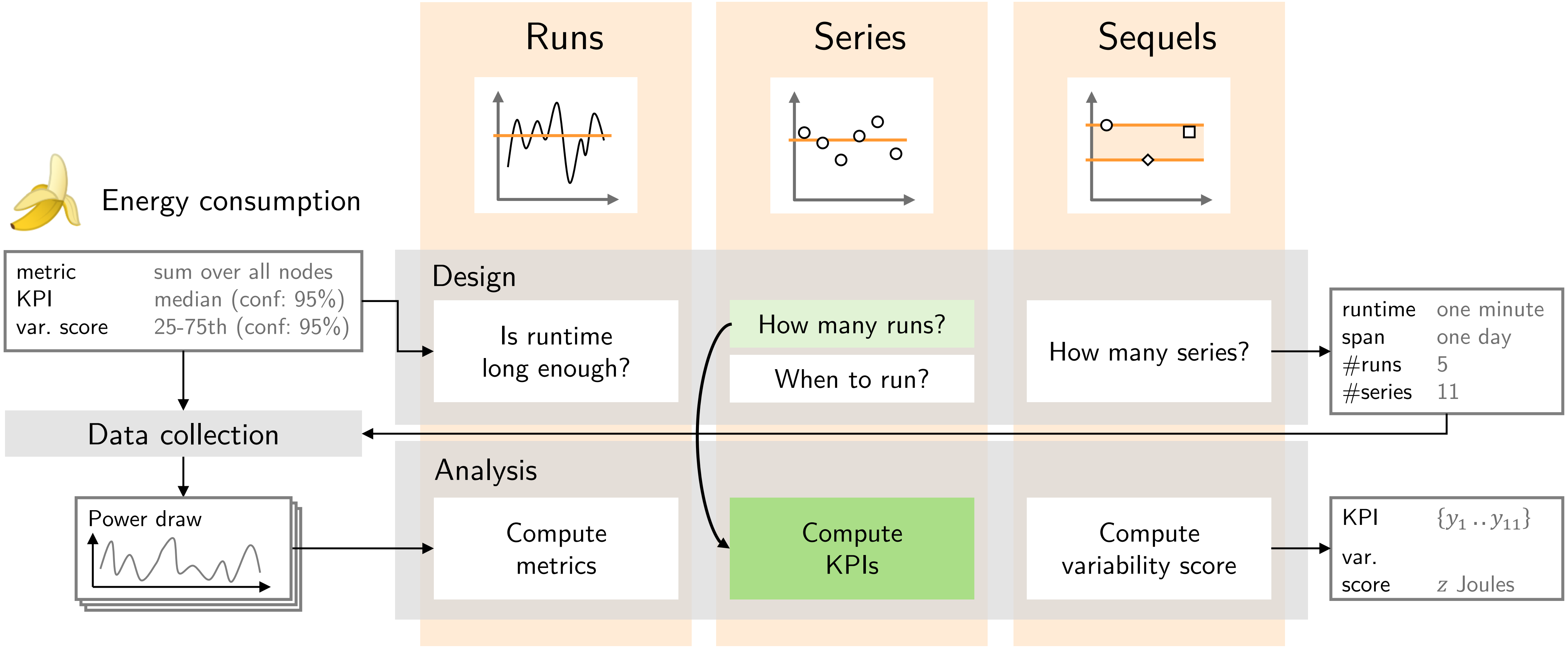
Enjoy your break!

45’ Lecture
10’ Hands-on

10’ Break

20’ Lecture

Wrap-up & Discussions


Runs
Series
Sequels
Energy consumption
metric sum over all nodes
KPI median (conf: 95%)
var. score 25-75th (conf: 95%)
Design
Is runtime long enough?
How many runs?
When to run?
How many series?
runtime one minute
span one day
#runs 5
#series 11
Data collection
Analysis
Power draw
Compute metrics
Compute KPIs
Compute variability score
KPI $\{y_1 \mathinner{..} y_{11}\}$
var. score z Joules

Assessing replicability

How to be fair and general?

Independence assumption

The elephant in the room

Dealing with seasonality

Patterns in networks

Assessing replicability

How to be fair and general?

Independence assumption

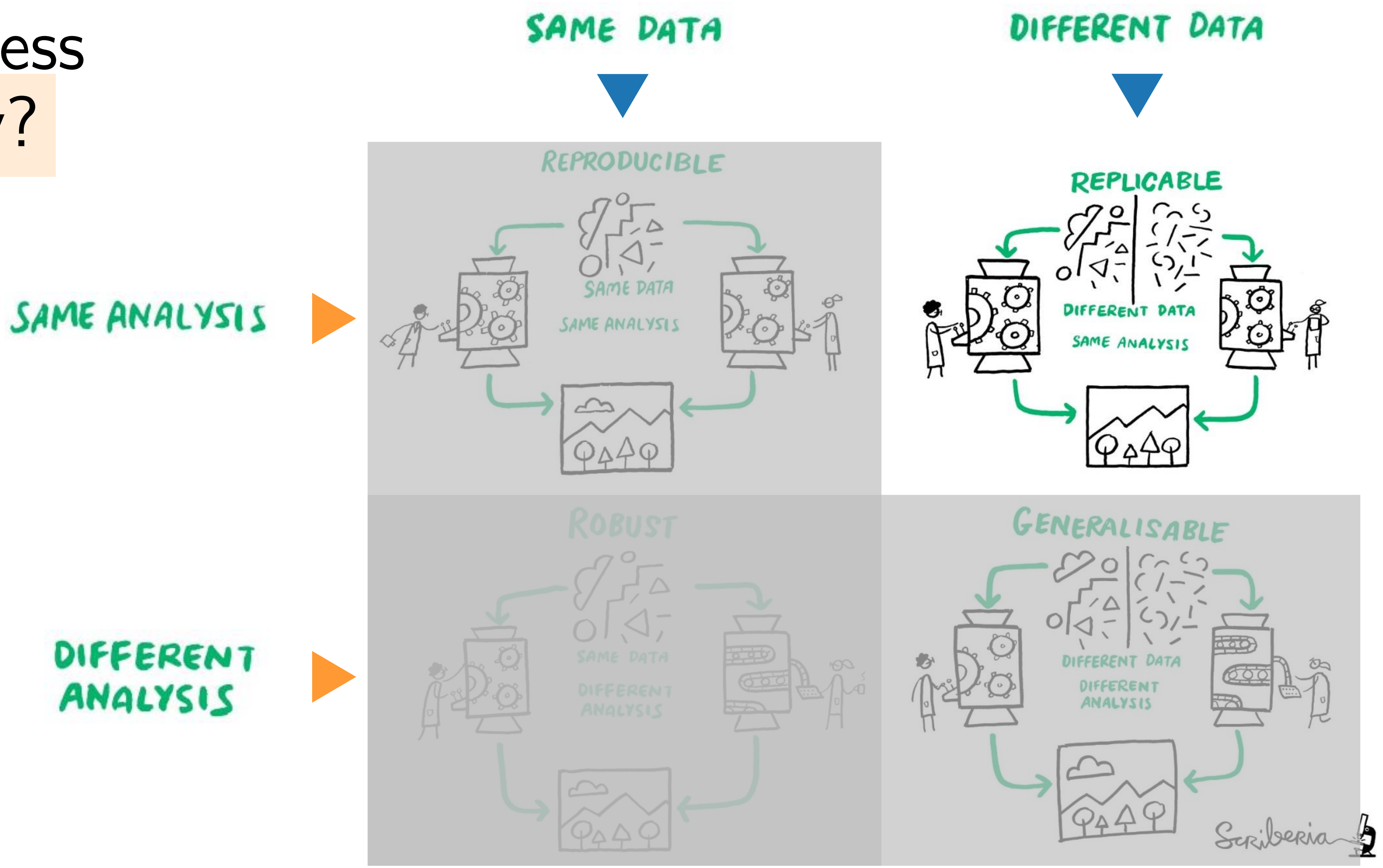
The elephant in the room

Dealing with seasonality

Patterns in networks

How to assess replicability?

What are "same" results?



The Turing Way project illustration by Scriberia. Zenodo. http://doi.org/10.5281/zenodo.3332807

How to assess replicability?

What are
"same" results?

Problems

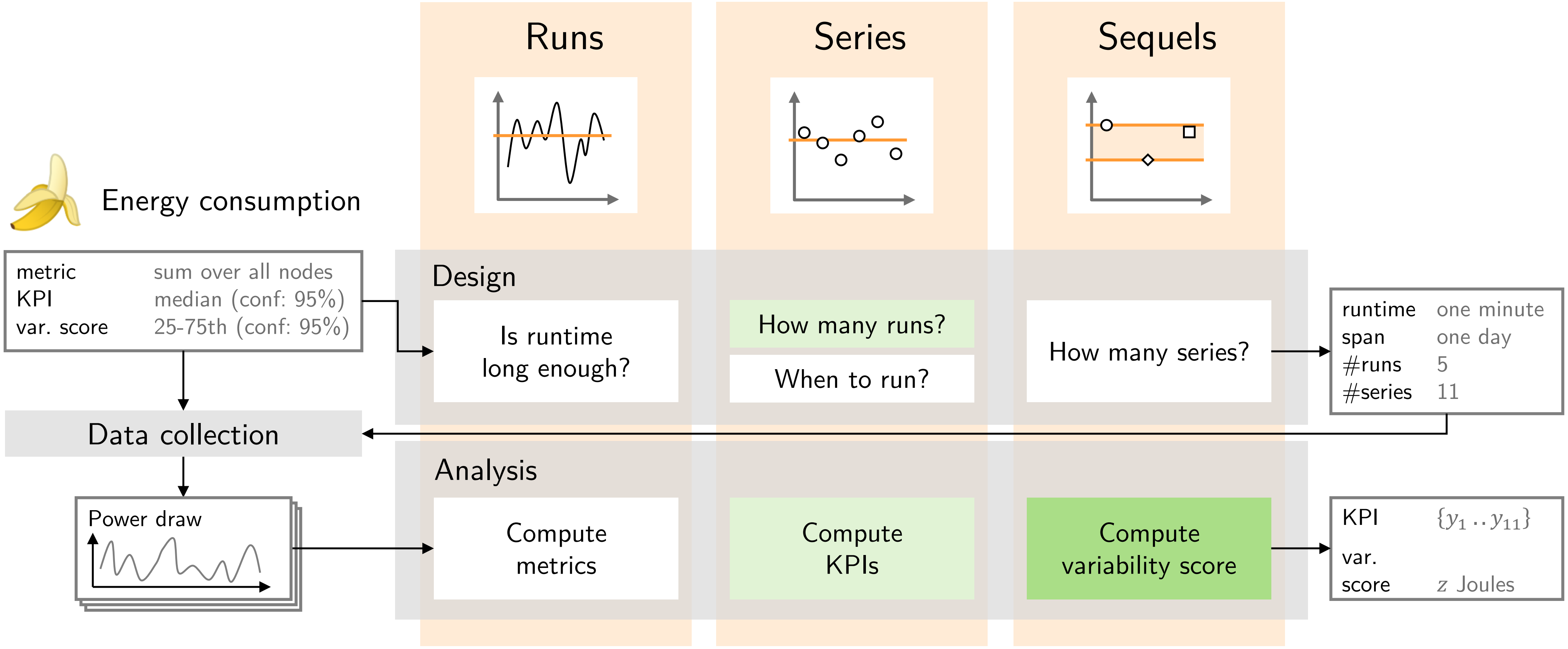
- Statistical tests are good at checking that things are different
- "Similarity" tests all boil down to testing whether differences below some threshold

Set how?

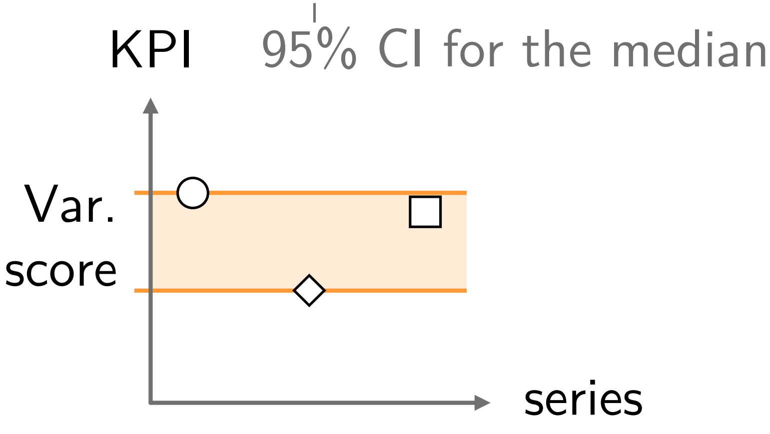
Our approach

Do not assess replicability as a binary criterion

Quantify variability


Runs
Series
Sequels
Energy consumption
metric sum over all nodes
KPI median (conf: 95%)
var. score 25-75th (conf: 95%)
Design
Is runtime long enough?
How many runs?
When to run?
How many series?
runtime one minute
span one day
#runs 5
#series 11
Data collection
Analysis
Power draw
Compute metrics
Compute KPIs
Compute variability score
KPI $\{y_1 \mathinner{..} y_{11}\}$
var. score z Joules

The Thompson's method provides non-parametric CI for distribution percentiles



Two-sided 75%-CI for the median

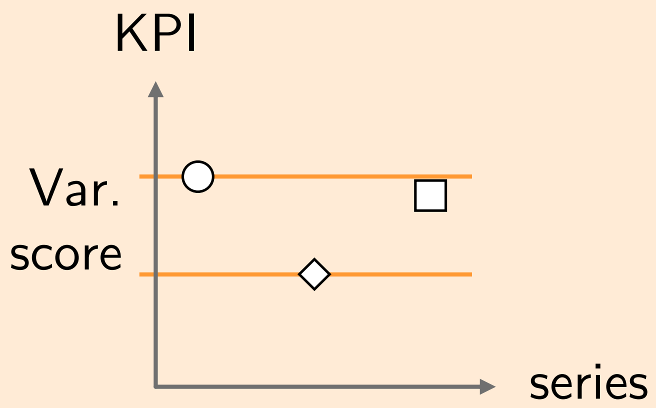
Variability scores are
percentile ranges
of KPI values

- Compute upper and lower bounds on the true percentile values for a certain confidence level

- Variability scores are defined as ranges between these bounds

The Thompson's method provides non-parametric CI for distribution percentiles

Variability scores are percentile ranges of KPI values

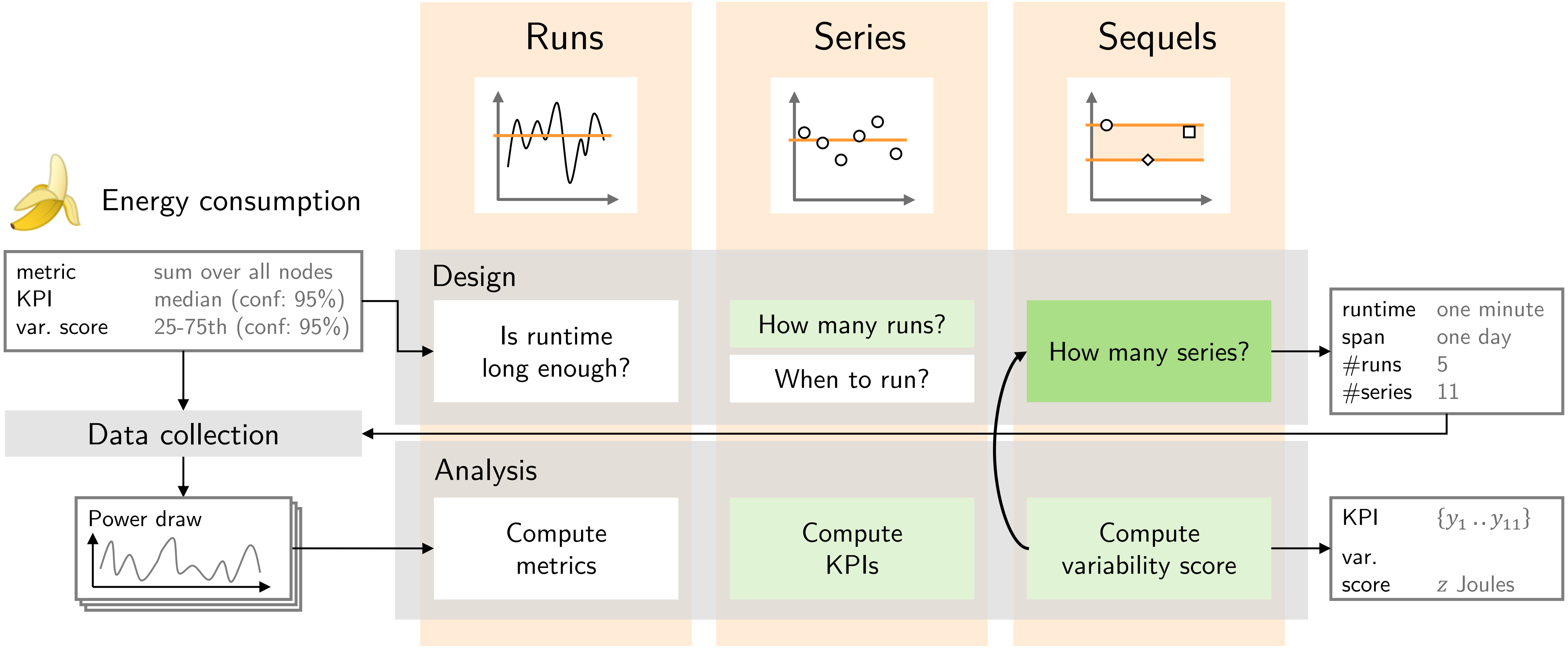


Two-sided 75%-CI for the median

- Compute upper and lower bounds on the true percentile values for a certain confidence level

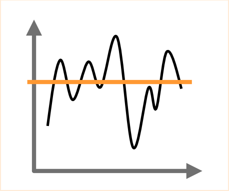
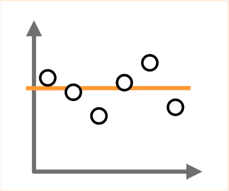
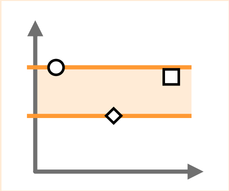
- Variability scores are defined as ranges between these bounds

▶ If a binary cut is desired, base it on the score


Runs
Series
Sequels
Energy consumption
metric
sum over all nodes
KPI
median (conf: 95%)
var. score
25-75th (conf: 95%)
Design
Is runtime long enough?
How many runs?
When to run?
How many series?
runtime
one minute
span
one day
#runs
5
#series
11
Data collection
Analysis
Power draw
Compute metrics
Compute KPIs
Compute variability score
KPI
$\{y_1 \ldots y_{11}\}$
var. score
z Joules

Let’s talk about

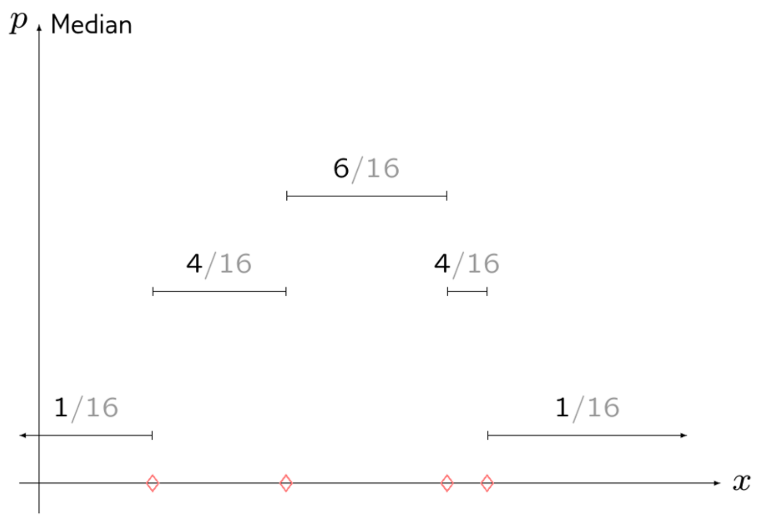
independence

	Runs	Series	Sequels
			
Design	Is runtime long enough?	How many runs? When to run?	How many series?
Analysis	Compute metrics	Compute KPIs	Compute variability score

Assessing replicability
How to be fair and general?

Independence assumption
The elephant in the room

Dealing with seasonality
Patterns in networks



Hypothesis

Samples are i.i.d.

I.I.D. is the acronym for Independent and Identically Distributed

I.I.D. is the acronym for Independent and Identically Distributed

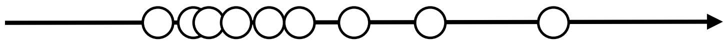
Memoriless i.e. Future samples are not correlated to past samples

I.I.D. is the acronym for Independent and Identically Distributed

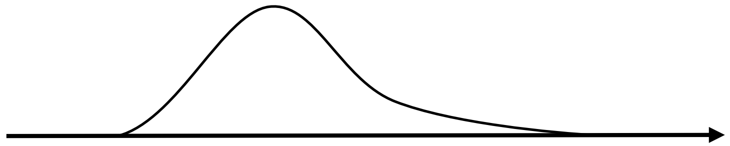
All samples are drawn from
the same underlying distribution

I.I.D. is the acronym for Independent and Identically Distributed

Identically distributed sample

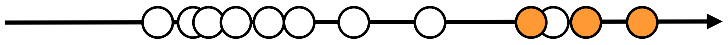


Underlying distribution

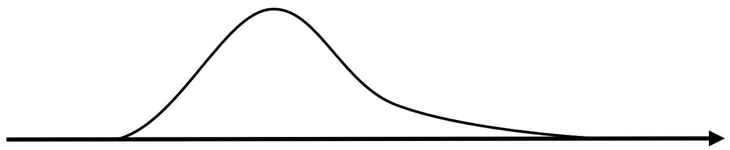


I.I.D. is the acronym for Independent and Identically Distributed

biased
~~Identically distributed~~ sample



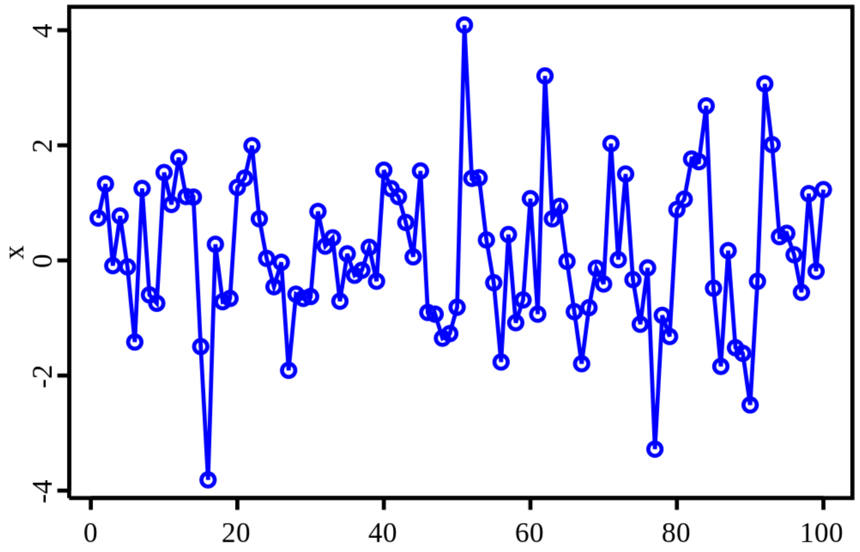
Underlying distribution
change



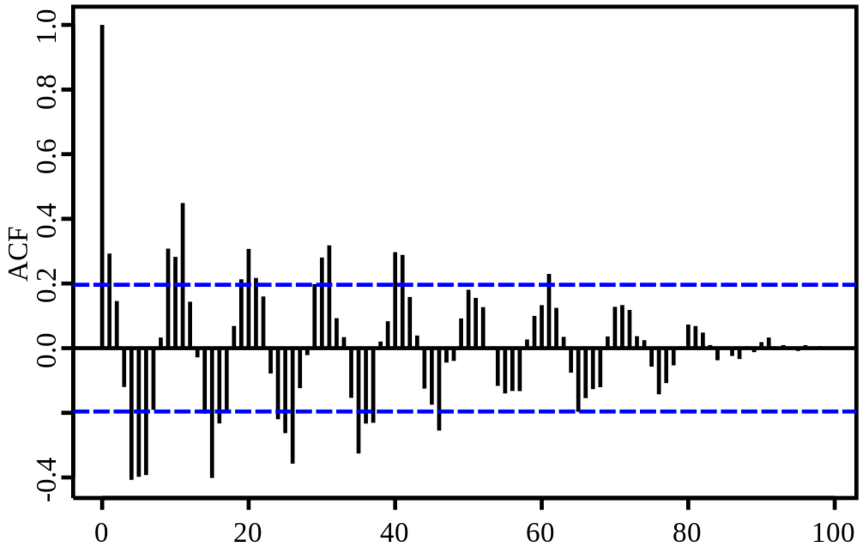
I.I.D. is the acronym for Independent and Identically Distributed

commons.wikimedia.org/wiki/File:Acf.svg

Correlation
i.e., non-independence
can be seen in an
autocorrelation plot



100 random numbers
with a "hidden" sine function



The autocorrelation plot reveals
the hidden structure in the data

In general

We often say "independence" when we mean "i.i.d.-ness"

What if there is no independence?

- Samples are biased
- Data do not contain as much information as it appears to.
- "Fake" effects

Independence is a property of the experiment design (not of the data!)

We often say or write

"Data is i.i.d."

Not mathematically correct statement

- We mean that the samples were collected from an i.i.d. experiment — ?

An experiment is i.i.d. if all its factors are selected in an i.i.d. way

Factor?

Any parameters that affect the outcome of an experiment

e.g., Time of the day

Factor values must be selected

- in a memoriless fashion — Independent
- using the same random procedure — Identically Distributed

An experiment is i.i.d. if all its factors are selected in an i.i.d. way, but this is often impossible

Uncontrollable factors	External interference may be unavoidable
Imperfect randomization	Experiments cannot overlap in time
Hidden factors	What about temperature?

Independence is often impossible to guarantee, but we can test if it appears to hold

Empirical i.i.d. test

- No trend
- No correlation structure

Implemented in TriScale

Two caveats

- Imprecise

  Especially with few samples

- No future guarantees

  Can only detect correlation that was captured in the sample

  ▶ Things may change...

Assessing replicability

How to be fair and general?

Independence assumption

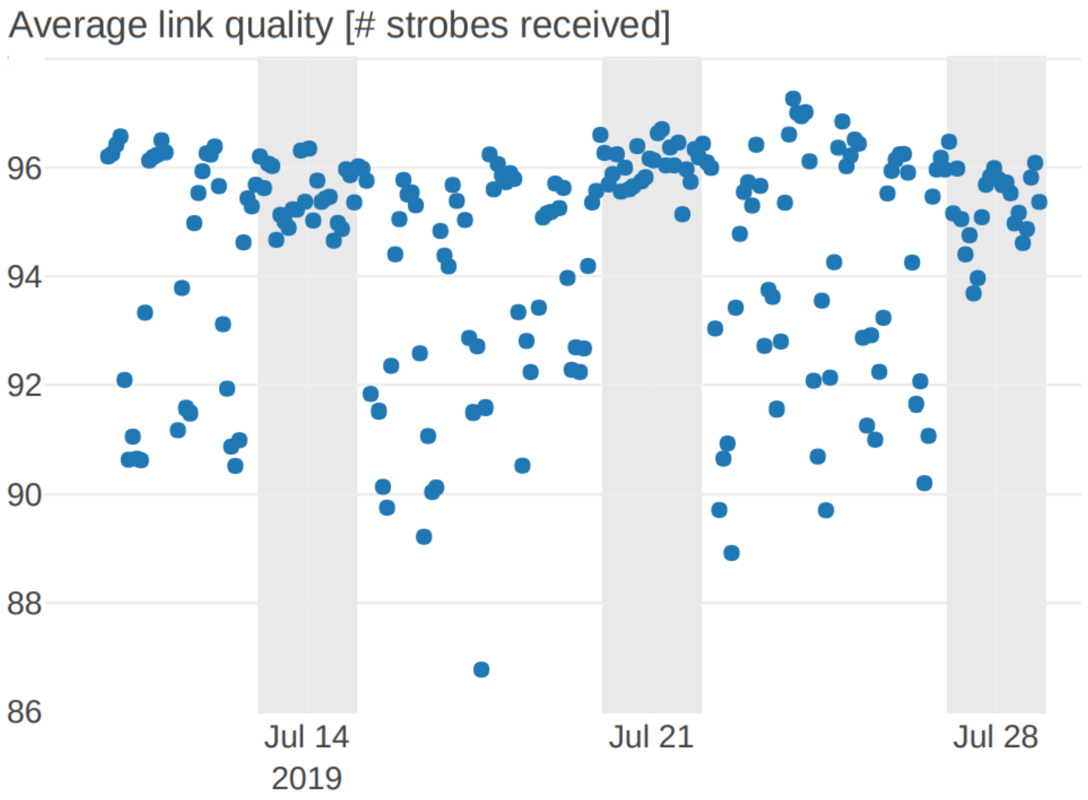
The elephant in the room

Dealing with seasonality

Patterns in networks

One common danger to beware of is seasonal components

Periodic patterns in the experimental conditions

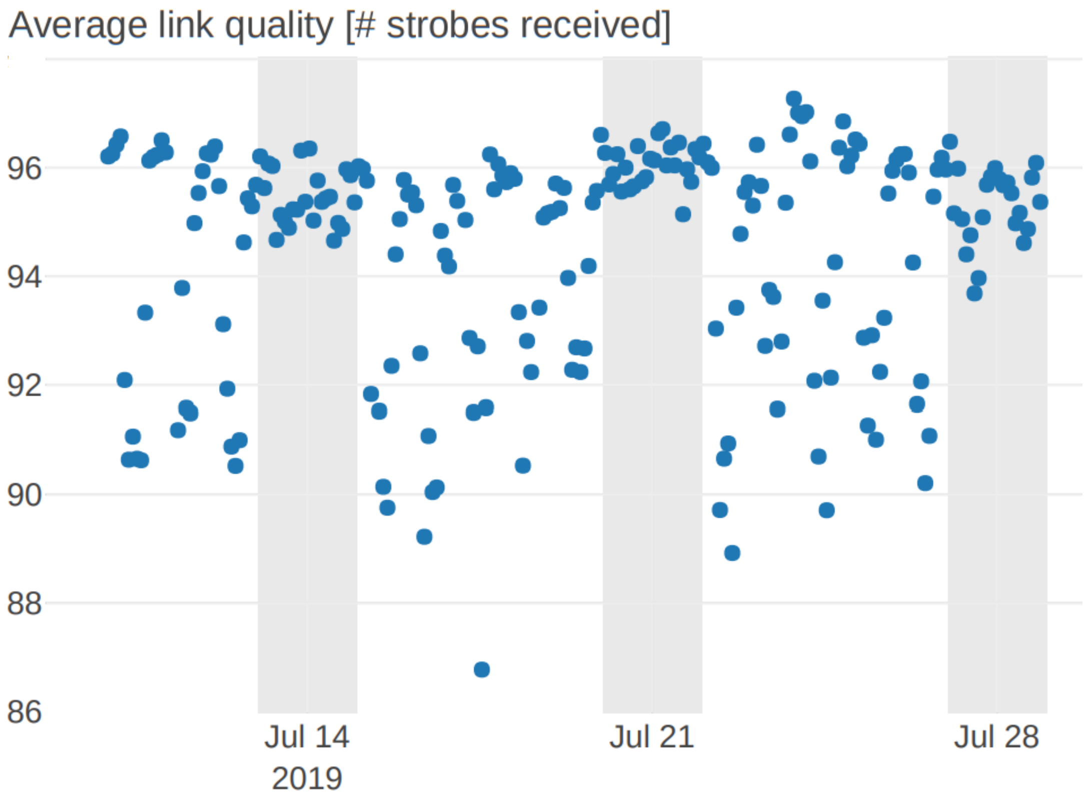


Average link quality on the Flocklab testbed
July 2019

One common danger to beware of is seasonal components

In TriScale

The time span of a series of runs should be a multiple of the largest seasonal component



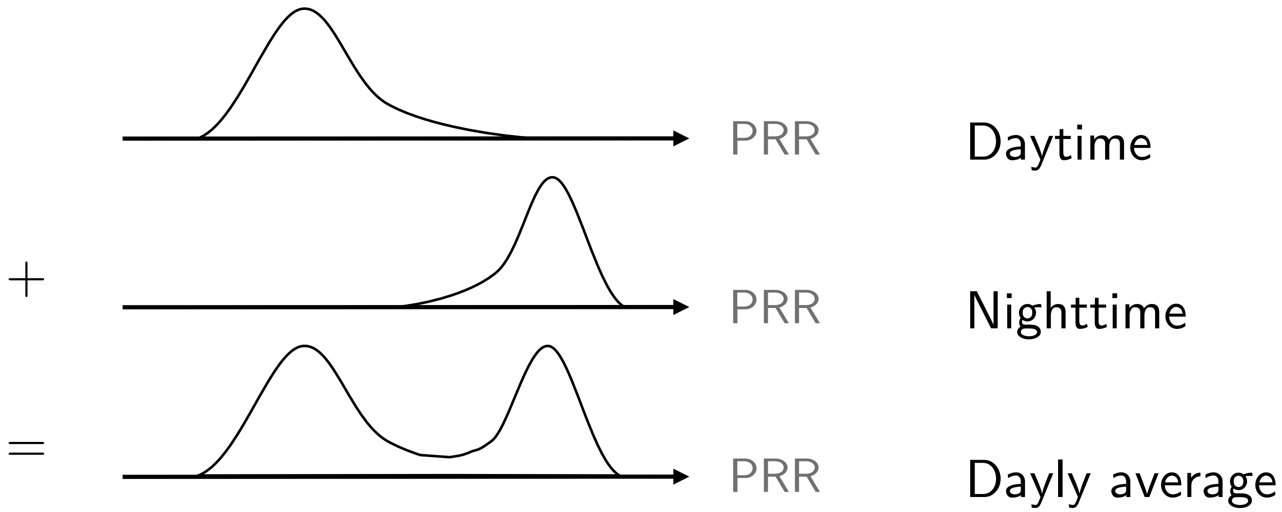
Average link quality on the Flocklab testbed
July 2019

One common danger to beware of is seasonal components

In TriScale

The time span of a series of runs should be a multiple of the largest seasonal component

Intuition



- Randomly sample this joint distribution
(not truly “identically distributed” experiment)

Identifying seasonnal components is a fairly difficult task

Requires

1. Long-term monitoring of the environment

2. Definition of a metric for “link quality” which is relevant for the system under test

   - Hidden factors!

Hard work but important!

We can see that in practice...

Any questions?

Up next

- Hands-on session
- Data analysis
- Seasonality

Let's practice!

Go to
triscale.ethz.ch

TriScale

TriScale

A Framework Supporting Replicable Performance Evaluations in Networking

View the Project on GitHub
romain-jacob/triscale

TriScale

A Framework Supporting Replicable Performance Evaluations in Networking

Paper Code Tutorial Discussion

Following a live tutorial session? Here are the links you're looking for

Hands-on	
Part 1	launch binder
Part 2	launch binder

When designing their performance evaluations, networking researchers often encounter questions such as:

- How long should a run be?
- How many runs to perform?
- How to account for the variability across multiple runs?
- What statistical methods should be used to analyze the data?

Despite the best intentions, researchers often answer these questions differently, thus impairing the replicability of evaluations and the confidence in the results.

Improving the standards of replicability has recently gained traction overall, as well as within the networking community. As an important piece of the puzzle, we developed a systematic methodology that streamlines the design and analysis of performance evaluations, and we have implemented this methodology into a framework called *TriScale*.

Why replicability matters
Case by example

Understanding variability
The three timescales

Know your data
Use the right statistics

Why replicability matters

Case by example

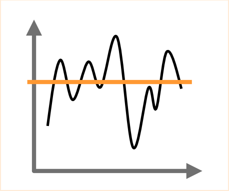
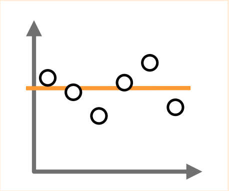
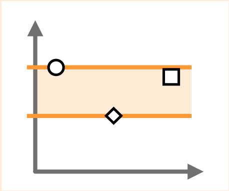
Understanding variability

The three timescales

Know your data

Use the right statistics

Some other boxes…

	Runs	Series	Sequels
			
Design	Is runtime long enough?	How many runs? When to run?	How many series?
Analysis	Compute metrics	Compute KPIs	Compute variability score

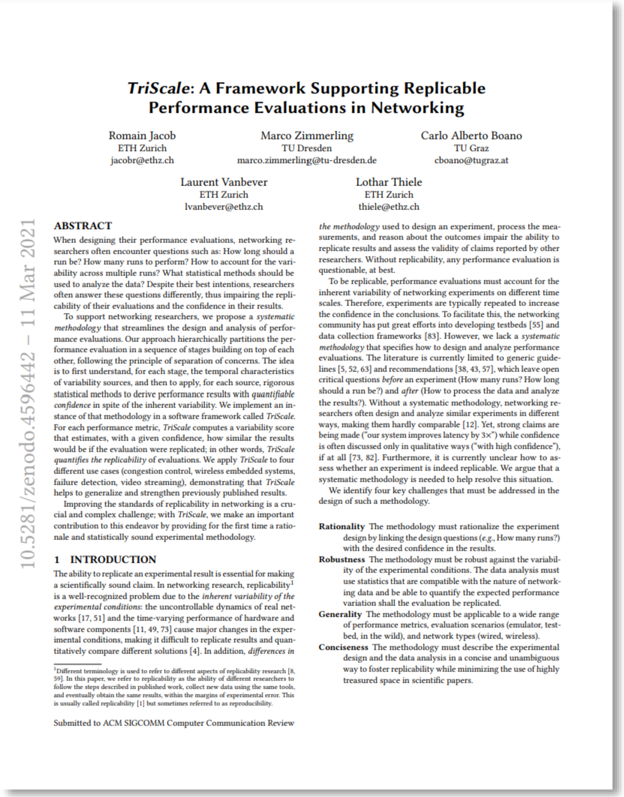
In TriScale

- Convergence of runs

For future work

- Comparison of confidence intervals

Interested?
Find our more!



TriScale: A Framework Supporting Replicable Performance Evaluations in Networking

Romain Jacob
ETH Zurich
jacobr@ethz.ch

Marco Zimmerling
TU Dresden
marco.zimmerling@tu-dresden.de

Carlo Alberto Boano
TU Graz
cboano@tugraz.at

Laurent Vanbever
ETH Zurich
lvanbever@ethz.ch

Lothar Thiele
ETH Zurich
thiele@ethz.ch

ABSTRACT

When designing their performance evaluations, networking researchers often encounter questions such as: How long should a run be? How many runs to perform? How to account for the variability across multiple runs? What statistical methods should be used to analyze the data? Despite their best intentions, researchers often answer these questions differently, thus impairing the replicability of their evaluations and the confidence in their results.

To support networking researchers, we propose a *systematic methodology* that streamlines the design and analysis of performance evaluations. Our approach hierarchically partitions the performance evaluation in a sequence of stages building on top of each other, following the principle of separation of concerns. The idea is to first understand, for each stage, the temporal characteristics of variability sources, and then to apply, for each source, rigorous statistical methods to derive performance results with *quantifiable confidence* in spite of the inherent variability. We implement an instance of that methodology in a software framework called *TriScale*. For each performance metric, *TriScale* computes a variability score that estimates, with a given confidence, how similar the results would be if the evaluation were replicated; in other words, *TriScale quantifies the replicability* of evaluations. We apply *TriScale* to four different use cases (congestion control, wireless embedded systems, failure detection, video streaming), demonstrating that *TriScale* helps to generalize and strengthen previously published results.

Improving the standards of replicability in networking is a crucial and complex challenge; with *TriScale*, we make an important contribution to this endeavor by providing for the first time a rationale and statistically sound experimental methodology.

1 INTRODUCTION

The ability to replicate an experimental result is essential for making a scientifically sound claim. In networking research, replicability[1] is a well-recognized problem due to the *inherent variability of the experimental conditions*: the uncontrollable dynamics of real networks [17, 51] and the time-varying performance of hardware and software components [11, 49, 73] cause major changes in the experimental conditions, making it difficult to replicate results and quantitatively compare different solutions [4]. In addition, *differences in the methodology* used to design an experiment, process the measurements, and reason about the outcomes impair the ability to replicate results and assess the validity of claims reported by other researchers. Without replicability, any performance evaluation is questionable, at best.

To be replicable, performance evaluations must account for the inherent variability of networking experiments on different time scales. Therefore, experiments are typically repeated to increase the confidence in the conclusions. To facilitate this, the networking community has put great efforts into developing testbeds [55] and data collection frameworks [83]. However, we lack a *systematic methodology* that specifies how to design and analyze performance evaluations. The literature is currently limited to generic guidelines [5, 52, 63] and recommendations [38, 43, 57], which leave open critical questions *before* an experiment (How many runs? How long should a run be?) and *after* (How to process the data and analyze the results?). Without a systematic methodology, networking researchers often design and analyze similar experiments in different ways, making them hardly comparable [12]. Yet, strong claims are being made ("our system improves latency by 3×") while confidence is often discussed only in qualitative ways ("with high confidence"), if at all [73, 82]. Furthermore, it is currently unclear how to assess whether an experiment is indeed replicable. We argue that a systematic methodology is needed to help resolve this situation.

We identify four key challenges that must be addressed in the design of such a methodology.

Rationality The methodology must rationalize the experiment design by linking the design questions (*e.g.*, How many runs?) with the desired confidence in the results.

Robustness The methodology must be robust against the variability of the experimental conditions. The data analysis must use statistics that are compatible with the nature of networking data and be able to quantify the expected performance variation shall the evaluation be replicated.

Generality The methodology must be applicable to a wide range of performance metrics, evaluation scenarios (emulator, testbed, in the wild), and network types (wired, wireless).

Conciseness The methodology must describe the experimental design and the data analysis in a concise and unambiguous way to foster replicability while minimizing the use of highly treasured space in scientific papers.

[1] Different terminology is used to refer to different aspects of replicability research [8, 59]. In this paper, we refer to replicability as the ability of different researchers to follow the steps described in published work, collect new data using the same tools, and eventually obtain the same results, within the margins of experimental error. This is usually called replicability [1] but sometimes referred to as reproducibility.

Submitted to ACM SIGCOMM Computer Communication Review

tiny.cc/triscale

triscale.ethz.ch

zenodo DOI 10.5281/zenodo.3464273

Getting the TriScale work published has been... complicated.

Rejected at

NSDI'20
SIGCOMM'20
SIGMETRICS'21
CCR'21

while receiving comments like

- Solid work with great tooling.
- Our community clearly has a problem with reproducibility and this paper presents very promissing solutions.
- Every PhD student should read this paper.

... wait what?

Getting the TriScale work published has been... complicated.

Rejected at	NSDI'20
	SIGCOMM'20
	SIGMETRICS'21
	CCR'21
Accepted at	JSys'21

while receiving comments like

- Solid work with great tooling.
- Our community clearly has a problem with reproducibility and this paper presents very promising solutions.
- Every PhD student should read this paper.

JSys

Journal of Systems Research

Diamond Open Access

Free to read. Free to publish.

No page limits

No paper caps

Different paper types

Solution	Usual ones
Problem	Position/white papers
SoK	Survey++
Tool	Typically hard to publish

Student board

Transparent reviewing

- Review summary
- Reviewers named
- Public reviews (anonym)

Artifact Evaluation

Independent but compulsory

JSys

Journal of Systems Research

13 different areas

- Networking
- Configuration Management for Systems
- Computer Architecture
- Real-time and Cyber-physical Systems
- Streaming Systems
- Systems for ML and ML for systems
- Distributed Consensus
- Data Science and Reproducibility
- Serverless Systems
- System Security
- Active Storage
- Wireless Embedded Systems
- Decentralized Systems

Journal of Systems Research

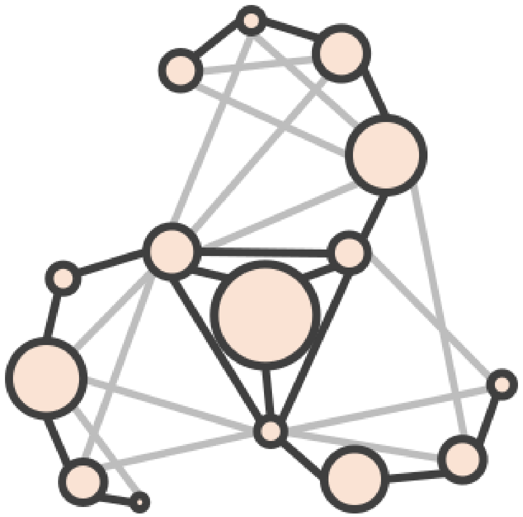
▶ jsys.org

13 different areas

- Networking

Area Chairs	Francis Yan Sangeetha Abdu Jyothi
Area Board	Amreesh Phokeer Ang Chen Arpit Gupta Colin Perkins Daehyeok Kim Srinivas Narayana

Experimental Reproducibility in Networking Research



Please fill out
our short survey!

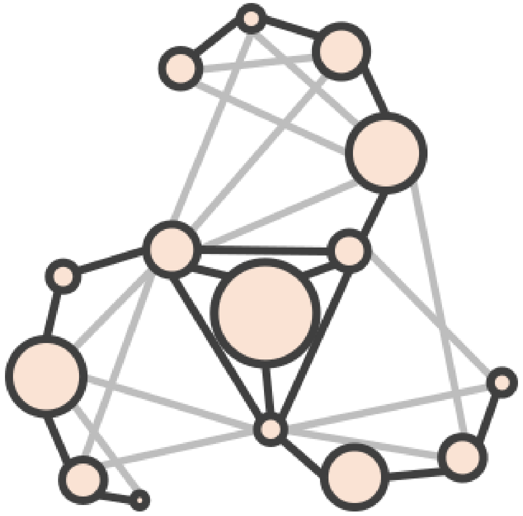
triscale.ethz.ch

feedback tutorial

Experimental Reproducibility in Networking Research

triscale.ethz.ch

feedback tutorial



Romain Jacob
ETH Zurich

Collaboration with
Marco Zimmerling, Laurent Vanbever
Carlo Alberto Boano, Lothar Thiele

@RJacobPartner