Assessment and Evaluation of Empirical and Scientific Data

Nikolaus Hansen
Inria & École polytechnique, France

November 2023

Why this talk?

- Explain to me why I am wrong (or miss something important)

The best state to be in is be either wrong or confused; 'cause it means there is more to learn.
— Laurence Krauss

- It’s a win either way (for me):
    - if you succeed, I learned something
    - otherwise,…

Why this talk?

- *If you are […] involved in a discussion or talking to an audience, ideally you should not try to persuade them, […]*
- *I am always put off by people who are called good speakers, by those who can arouse an audience. That's just what you do not want. If you have the capacity to do it, you should suppress it.*
- *rhetoric is the art [...] of persuading people by appealing to their emotions, [...] undermining their capacity for independent though and inquiry […] it's exactly the opposite of what it ought to be.*

— Noam Chomsky

Why this talk?

elevate your capacity for independent thought and inquire

Logic of (empirical) Research

- in mathematical logic, ***universal* statements** (“for all”, ∀) are *not derivable* from singular statements (“there exists”, ∃), however, universal statements **can be *contradicted* by singular statements** ($\Longrightarrow$ falsifiability)

- **single occurrences are of no significance** to science, science aims to make universal claims

single occurrences imply “there exists”, a scientific statement should read “for all”…

$\Longrightarrow$ replicability is the criterion of demarcation

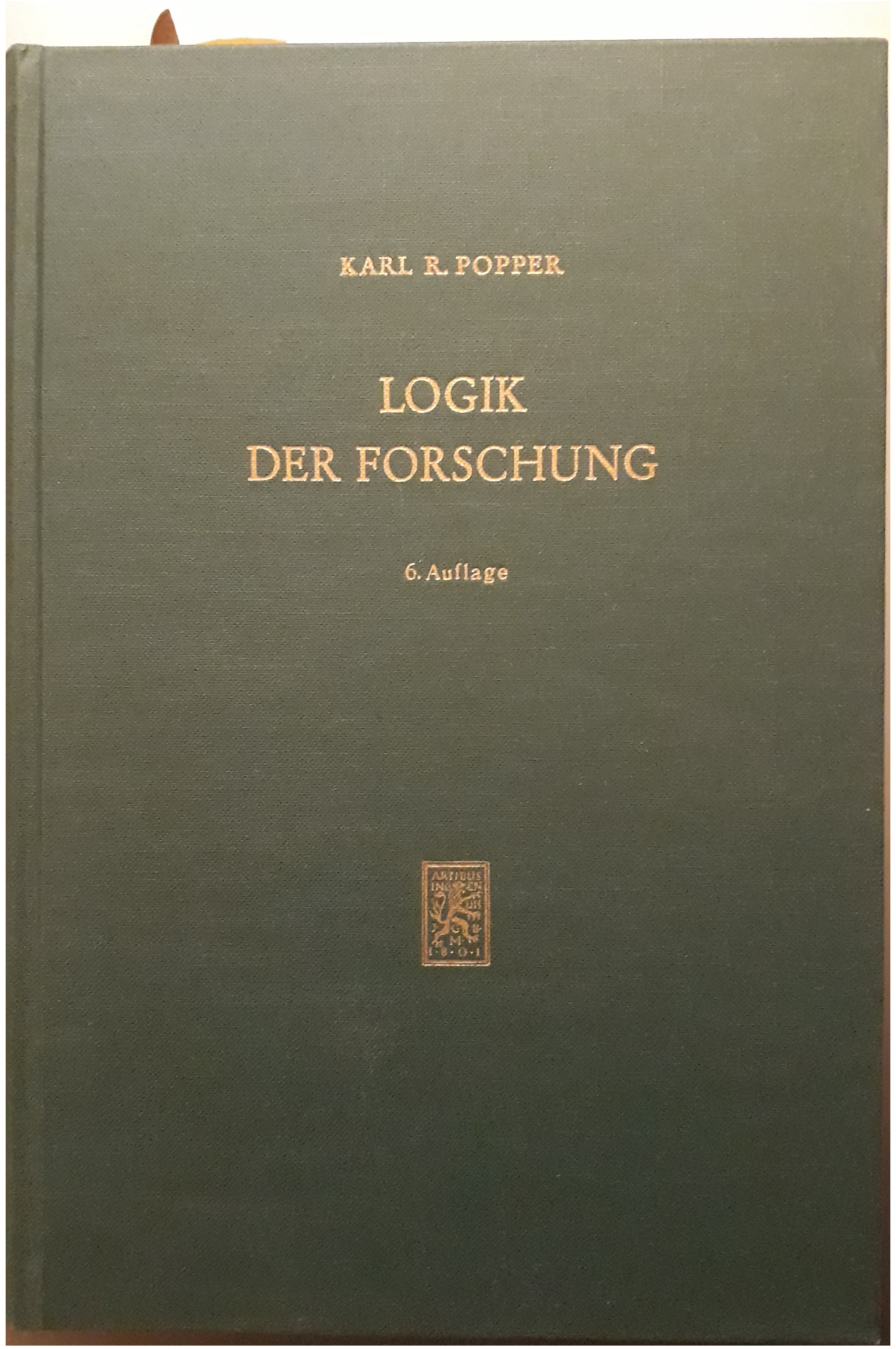
$\Longrightarrow$ universal and falsifiable (by failed replications)

$\Longrightarrow$ intrinsically incremental (many people succeed to replicate)

not only standing on the shoulders of giants

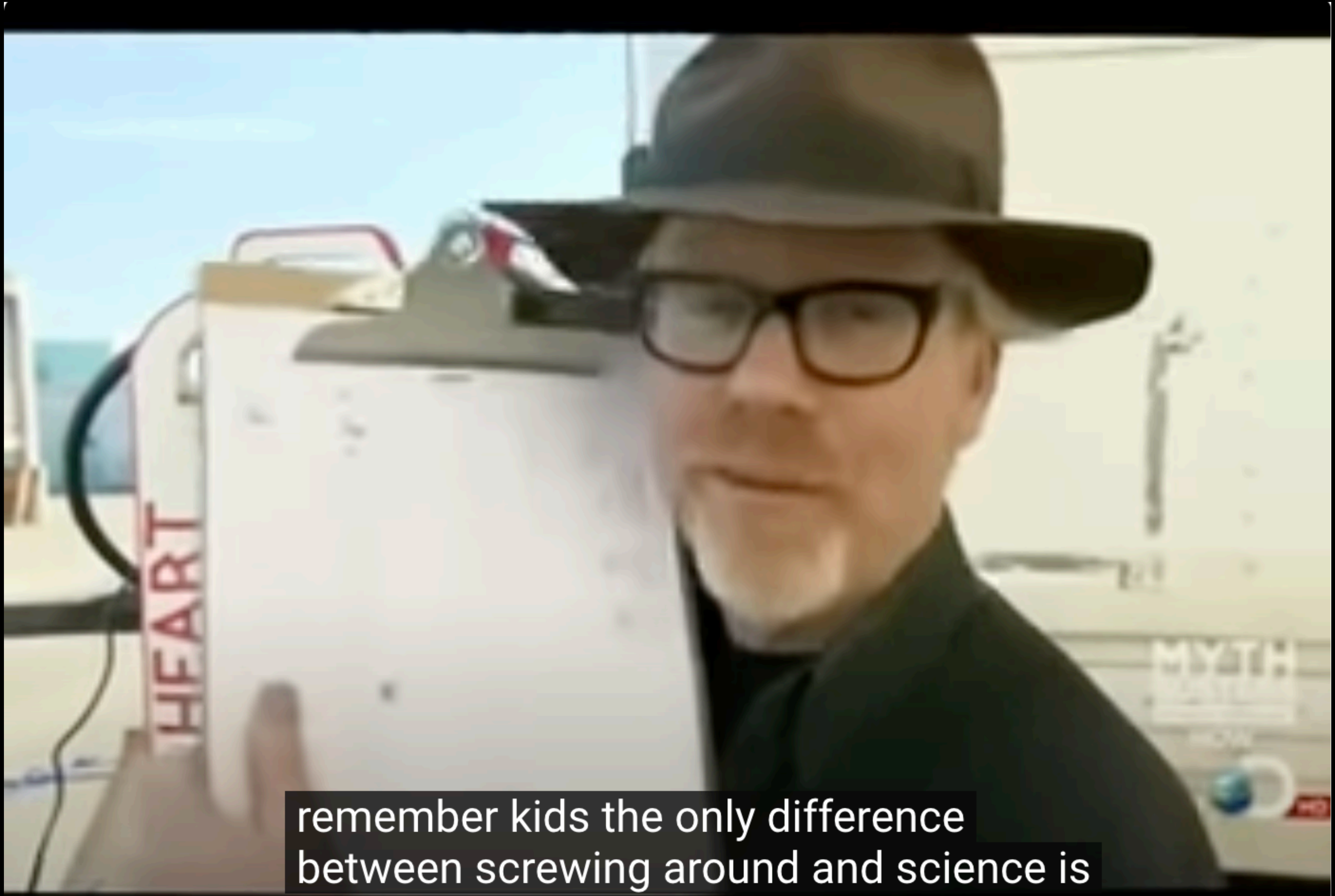
$\Longrightarrow$ consensus (“undeniable” wisdom)

$\Longrightarrow$ knowledge



Popper 1959. The logic of scientific discovery

Nikolaus Hansen, Inria



— Adam Savage (on YouTube)

writing it down

Levels of Reproducibility

Reproducibility in Evolutionary Computation 14:7

Table 1. Proposed Classification of Reproducibility Studies

Label	Artifacts	Random factors	Fixed factors	Purpose of the study
Repeatability	Original	Original	Original	Exactly repeat the original experiment, generating precisely the same results.
Reproducibility	Original	New	Original	Test whether the original results were dependent on specific values of random factors and, hence, only a statistical anomaly.
Replicability	New	New	Original	Test whether it is possible to independently reach the same conclusion without relying on original artifacts.
Generalisability	Original or New	New	New	Test whether the conclusion extends beyond the experimental setup of the original paper. When new artifacts are used, generalisability should come after a replicability study.

Source: López-Ibáñez et al. 2021. *ACM TELO* 1, 4.

corroboration: different result that is consistent with or supportive of the original conclusion (an *inconsistent* result could falsify the original claim)

Why are papers not replicable?

Ioannidis 2005. Why most published research findings are false. *PLoS medicine*, 2(8).
Bishop 2019. Rein in the four horsemen of irreproducibility. *Nature*, 568(7753).
Cockburn et al. 2020. Threats of a replication crisis in empirical computer science. *Communications of the ACM*, 63(8).

The obvious:

- small study size

  always do a second run

- bugs or trivial oversights

  particularly common for comparison/competitor algorithms

- selective/distorted/misleading/exaggerating reporting
- outright fraud

The less obvious:

- small effect size

  $\Longrightarrow$ false negative or misleading positive outcome

- misinterpreting the p-value:
  - high "multiplicity" $\Longrightarrow$ selection and publication bias (we don't report failures)
    - great number of tested relationships (p-fishing)
    - many teams independently working in parallel
    - great flexibility in study design and analytical modes (methods)
  - small ratio of true to false hypotheses (Odds$(H_0) \gg 1$, a good algorithm is difficult to improve)
  - confusion between hypothesis generating (HARKing) data and hypothesis testing (evidential) data

concerns authors and readers

Why are papers not replicable?

Ioannidis 2005. Why most published research findings are false. *PLoS medicine*, 2(8).
Bishop 2019. Rein in the four horsemen of irreproducibility. *Nature*, 568(7753).
Cockburn et al. 2020. Threats of a replication crisis in empirical computer science. *Communications of the ACM*, 63(8).

The obvious:

- small study size

  always do a second run

- bugs or trivial oversights

  particularly common for comparison/competitor algorithms

- selective/distorted/misleading/exaggerating reporting
- outright fraud

The less obvious:

- small effect size

  $\Longrightarrow$ false negative or misleading positive outcome

- misinterpreting the p-value:
  - high "multiplicity" $\Longrightarrow$ selection and publication bias (we don't report failures)
    - great number of tested relationships (p-fishing)
    - many teams independently working in parallel
    - great flexibility in study design and analytical modes (methods)
  - small ratio of true to false hypotheses (Odds$(H_0) \gg 1$, a good algorithm is difficult to improve)
  - confusion between hypothesis generating (HARKing) data and hypothesis testing (evidential) data

concerns authors and readers

Why are papers not replicable?

The less obvious:

- small effect size

$\Longrightarrow$ false negative or misleading positive outcome

- misinterpreting the p-value:
  - high "multiplicity" $\Longrightarrow$ selection and publication bias (we don't report failures)
    - great number of tested relationships (p-fishing)
    - many teams independently working in parallel
    - great flexibility in study design and analytical modes (methods)
  - small ratio of true to false hypotheses (Odds$(H_0) \gg 1$, a good algorithm is difficult to improve)
  - confusion between hypothesis generating (HARKing) data and hypothesis testing (evidential) data

concerns authors and readers

Not all of these points are *in itself* a problem and some of them are *intrinsic* to the process.

that is, even a "flawless" paper can be a statistical fluke!

The "reproducibility crisis" may come as much from the *interpretation* of scientific literature as from its *production*.
We often "forget" that science is incremental by construction.

in particular, interpreting hypothesis-*generating* publications as hypothesis-*confirming*
interpreting the p-value to mean $P(H_0 \mid D)$
considering a peer-reviewed paper as ground truth

Nikolaus Hansen, Inria

How to test replicability?

Q: What is the best evidence for the claim that a paper is replicable?

A: The paper *has been* replicated (the more often the better)

by independently peer-reviewed papers (ideally from different authors)
that are *crucially based on* the result in question

Colquhoun 2019: "*In the end, **the only way to solve the problem of reproducibility is to do more replication** and to reduce the incentives that are imposed on scientists to produce unreliable work.*"

The False Positive Risk: A Proposal Concerning What to Do About *p*-Values.
The American Statistician, 73:sub1.

Instead of "replicability" as a categorical true-or-false statement, consider the *probability* that a paper is in essence correct (replicable) by using all currently available evidence.

Quantification

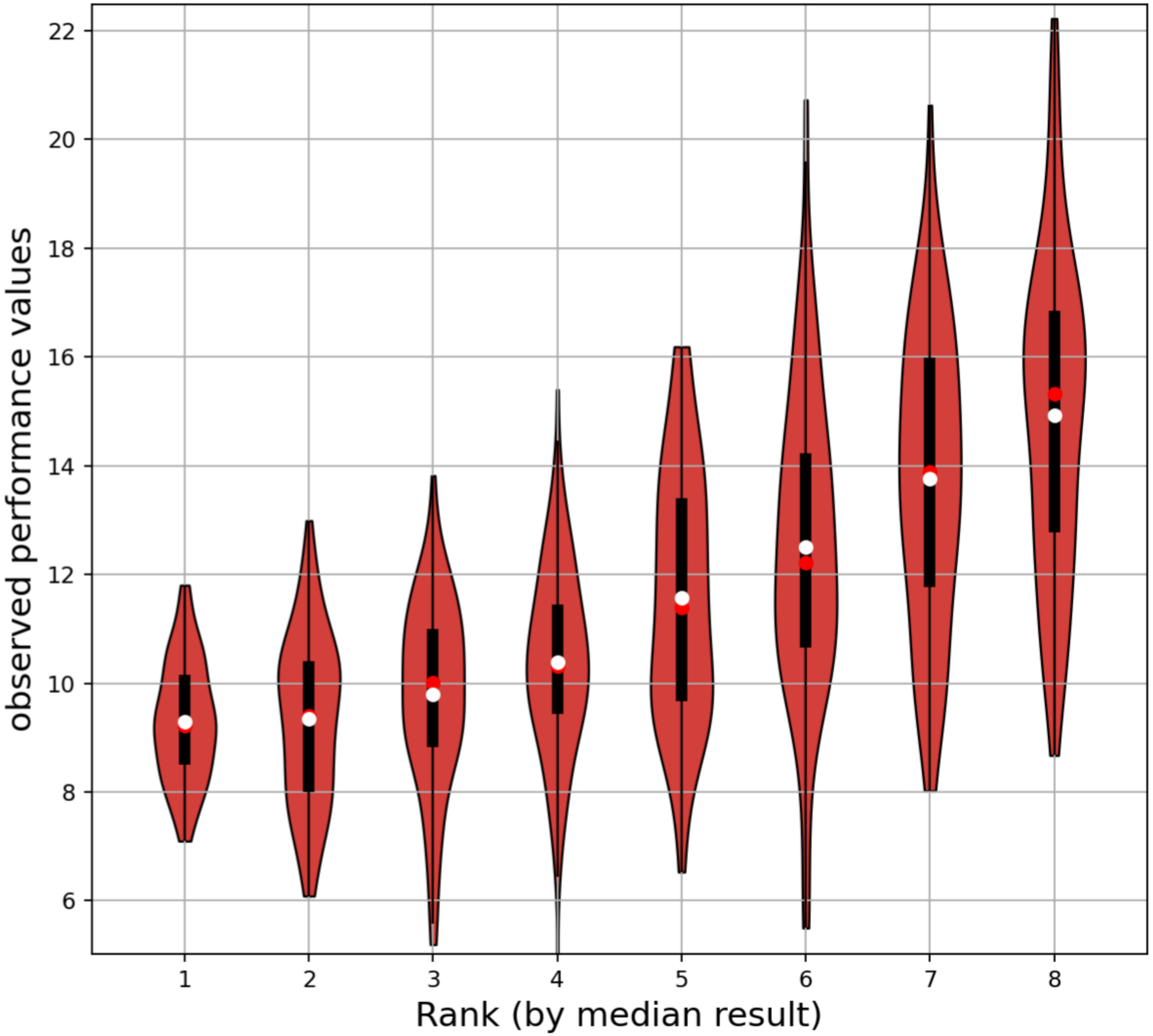
quantify quantify quantify

Quantification

Quantify.

Sagan 1995. The demon haunted world. 12: The fine art of baloney detection.

What's wrong with ranking algorithms?



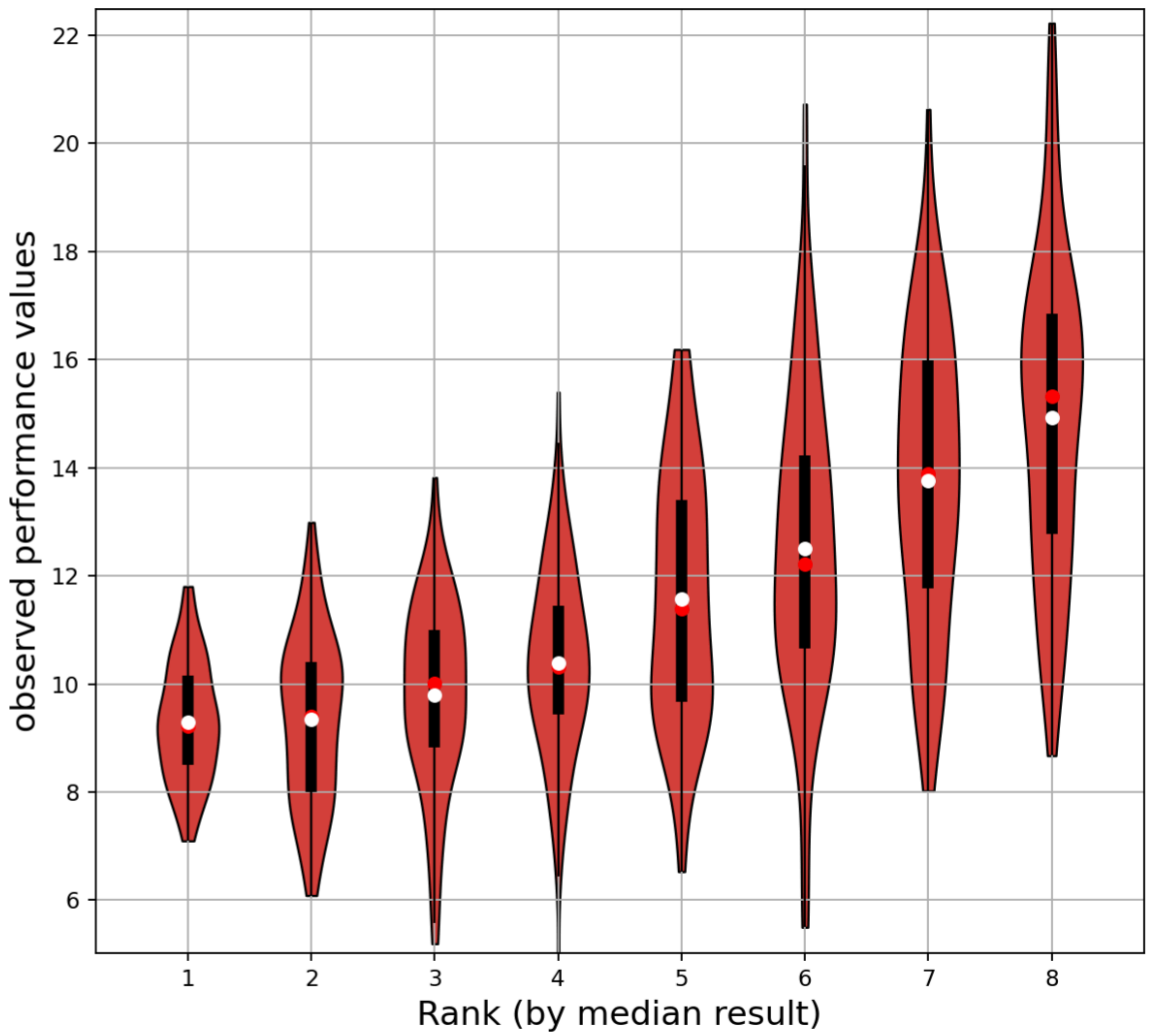
Inspired by Hoos 2023. Inconvenient Truths on Algorithm Competitions and Ways of Improving on Known Weaknesses. Presented at the Dagstuhl seminar 23251 *Challenges in Benchmarking Optimization Heuristics*.

What's wrong with ranking algorithms?

The ranking

- erases the information about effect size, hence relevance of a rank difference

- lacks a consistent distinction between (genuinely) equal and non-equal ranks



a mutual tie of algorithm pairs (1, 2), (2, 3) and (3, 4) does not imply a tie of (1, 4)

Nikolaus Hansen, Inria

Four Levels of Measurement (Scales)

- Nominal - categorial, define a classification
- Ordinal - define an order, e.g., **ranks**, function values (arguably)
- Interval - differences are quantitatively meaningful
- Ratio - ratios are meaningful, has a true zero, we can take the logarithm, e.g., time, function evaluations, iterations, odds, p-values

Stevens 1946. On the Theory of Scales of Measurement. *Science,* 103 (2684).

Measuring Performance

A performance measure should ideally be

- quantitative on the ratio level (highest level of measurement)

  logarithms are meaningful for assessing order of magnitudes
  "algorithm A is two *times* better than algorithm B"
  as "performance(B) / performance(A) = 1/2 = 0.5"
  should be semantically meaningful statements

- assuming a wide range of values

- comparable between different algorithms and across publications

- meaningfully interpretable and relevant (in the real world)

Runtime is a prime example when measured in an easily reproducible unit (evaluations, iterations, episodes).

Hansen et al. 2022. Anytime performance assessment in blackbox optimization benchmarking. *IEEE Trans. on EC*, 26(6).

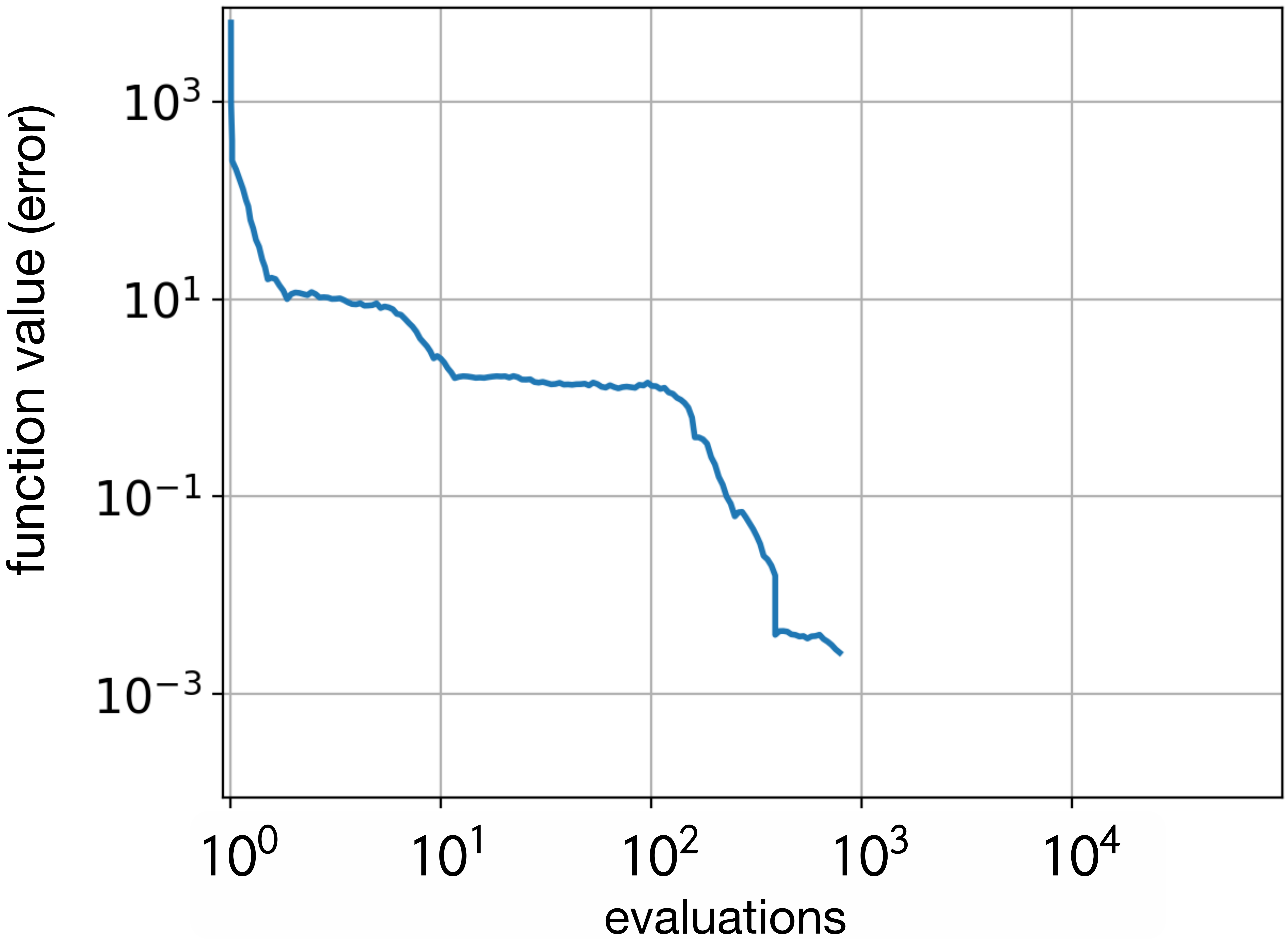
Empirical Cumulative Distributions

Empirical cumulative distribution functions (ECDF, or in short, *empirical distributions*) are arguably the most powerful tool to collect ("aggregate") many data points of the same unit of measurement in a single graph.

Main technique used in the COCO benchmarking platform.

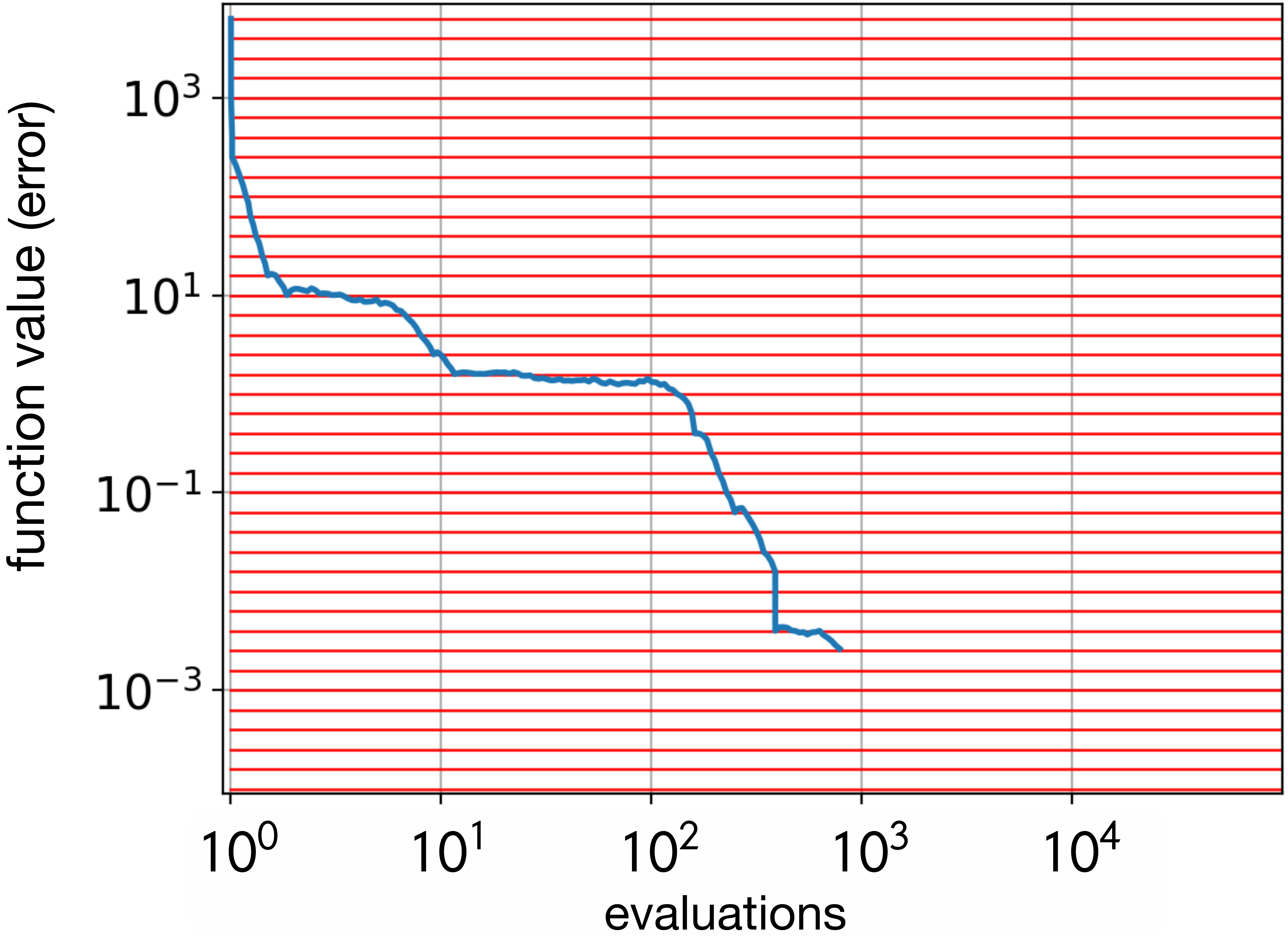
Hansen et al. 2021. A platform for comparing continuous optimizers in a black-box setting. *Optimization Methods and Software*, *36*(1).

From a Convergence Graph to the Empirical Runtime Distribution



Hansen et al. 2022. Anytime performance assessment in blackbox optimization benchmarking. *IEEE Trans. on EC*, 26(6).

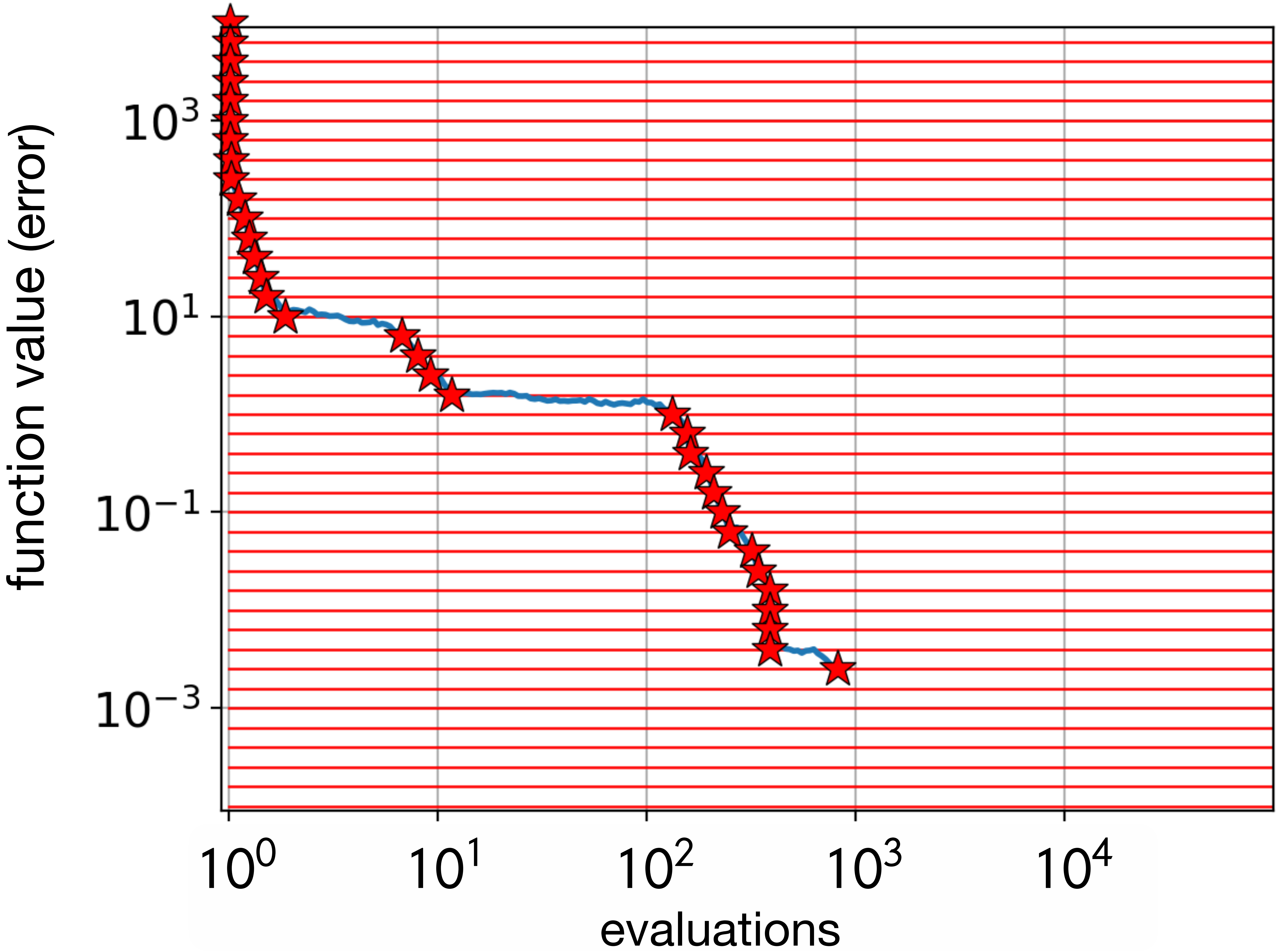
From a Convergence Graph to the Empirical Runtime Distribution



equidistance
“target” values

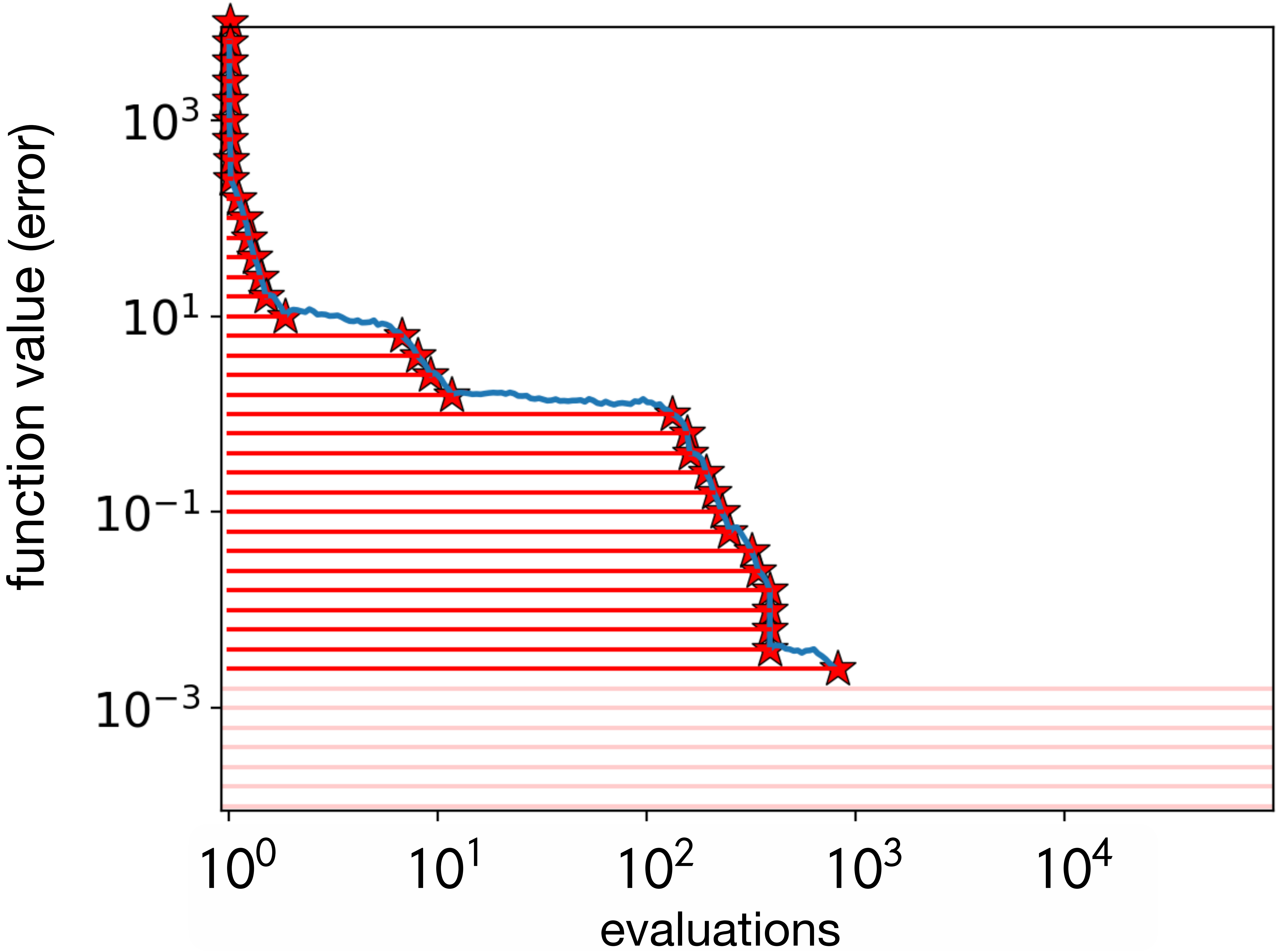
Hansen et al. 2022. Anytime performance assessment in blackbox optimization benchmarking. *IEEE Trans. on EC*, 26(6).

From a Convergence Graph to the Empirical Runtime Distribution



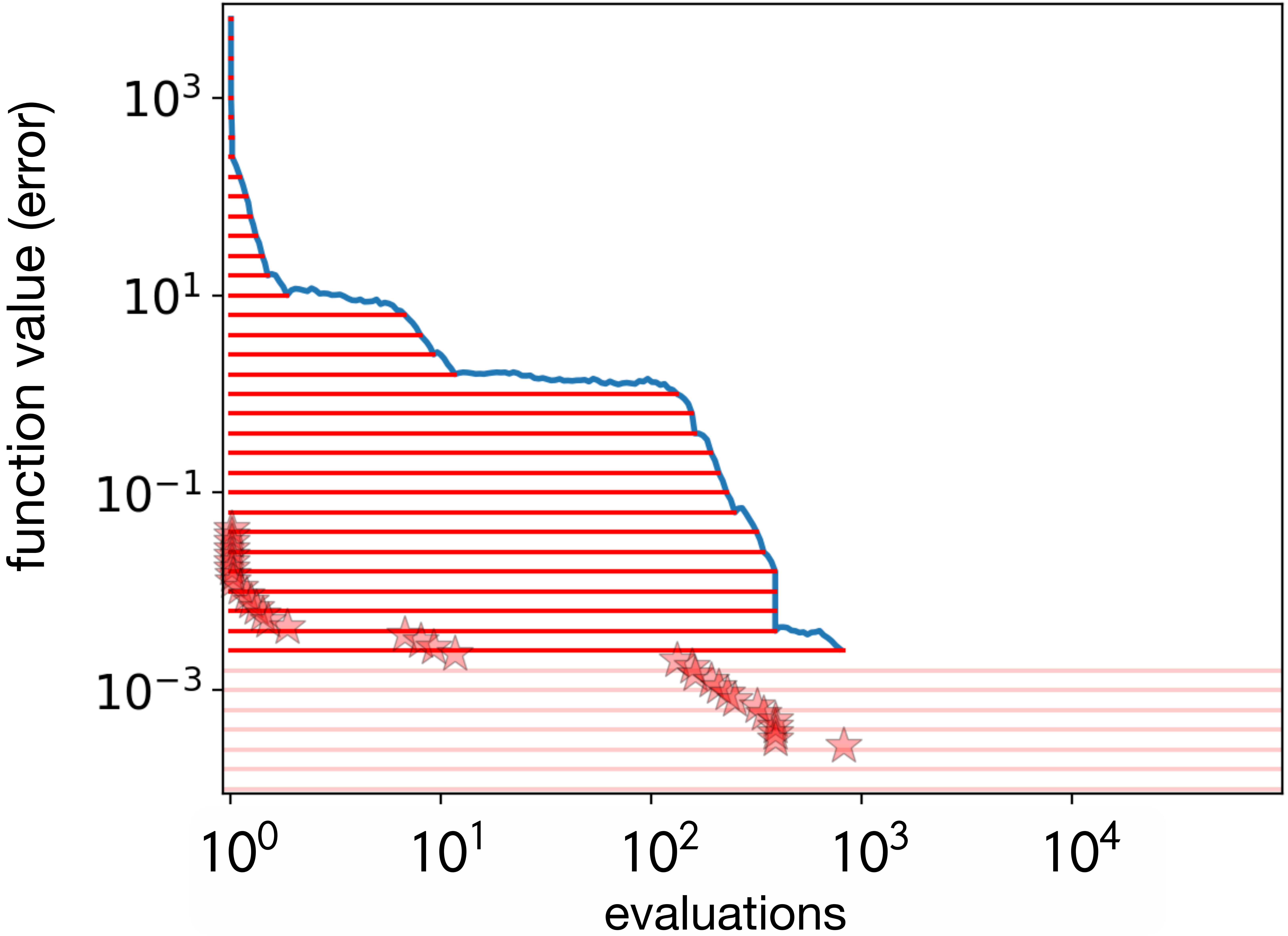
Hansen et al. 2022. Anytime performance assessment in blackbox optimization benchmarking. *IEEE Trans. on EC*, 26(6).

From a Convergence Graph to the Empirical Runtime Distribution



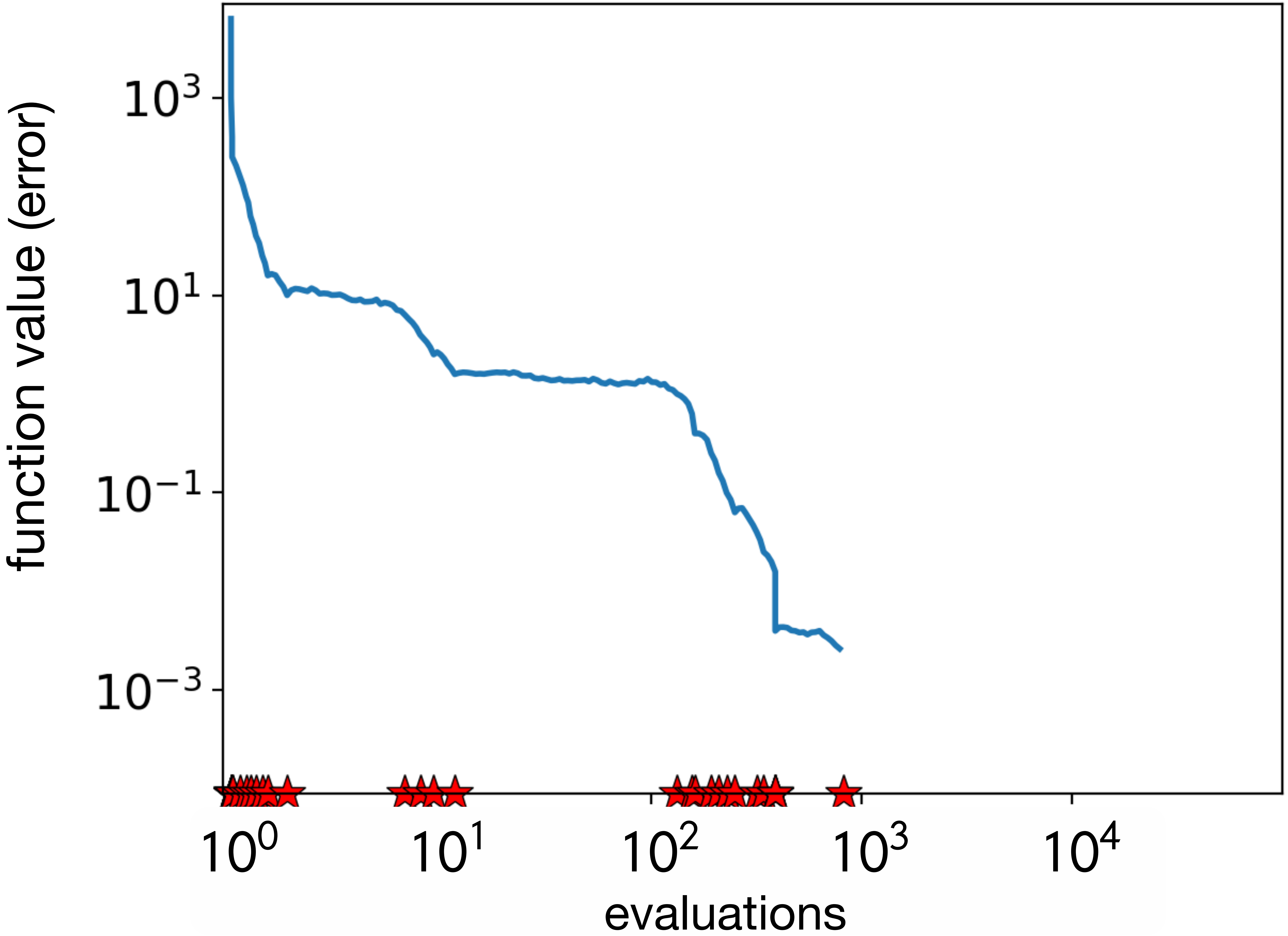
Hansen et al. 2022. Anytime performance assessment in blackbox optimization benchmarking. *IEEE Trans. on EC*, 26(6).

From a Convergence Graph to the Empirical Runtime Distribution



Hansen et al. 2022. Anytime performance assessment in blackbox optimization benchmarking. *IEEE Trans. on EC*, 26(6).

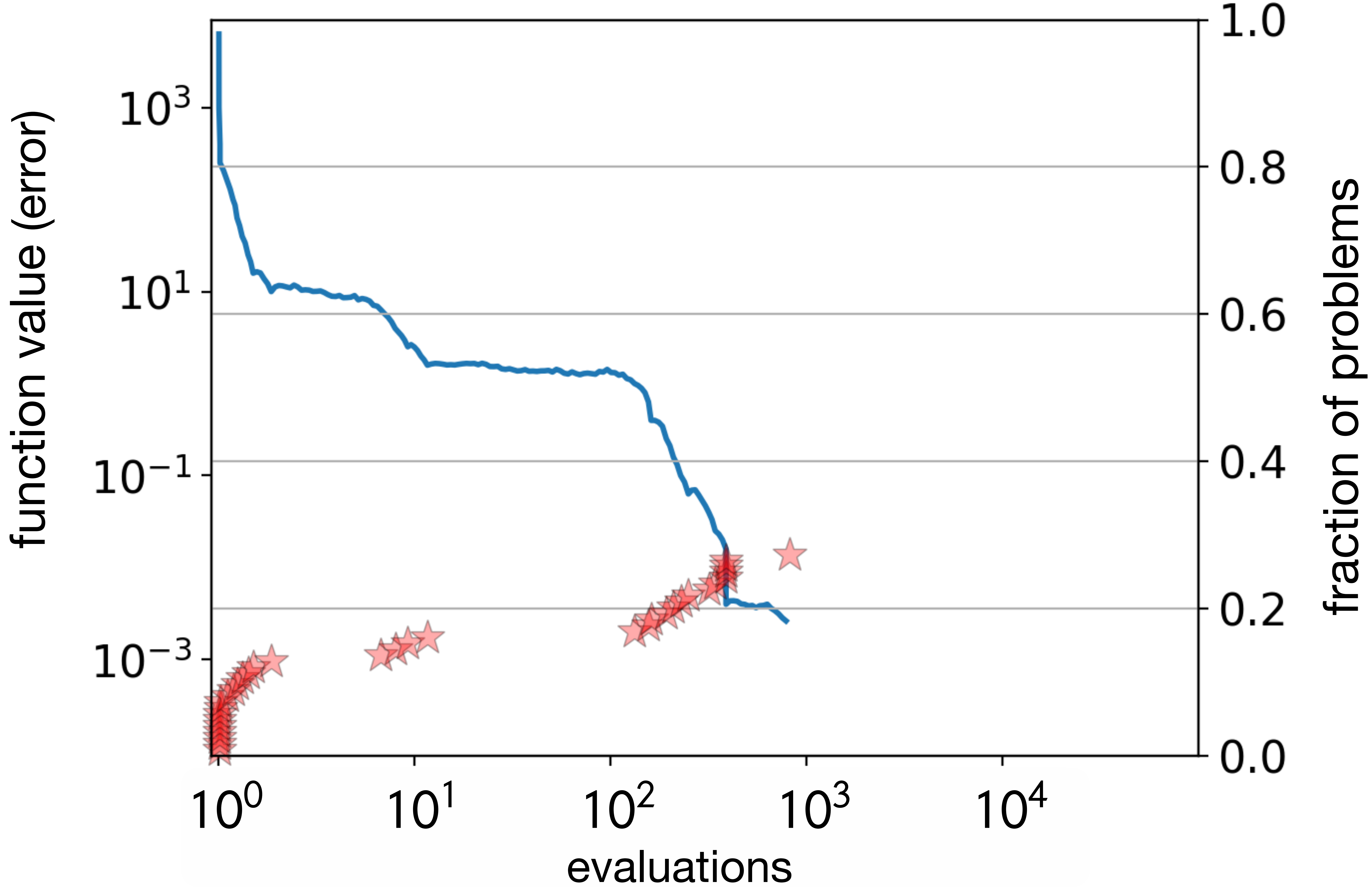
From a Convergence Graph to the Empirical Runtime Distribution



for the remaining construction, we could use any runtimes, for example, from different runs or even functions

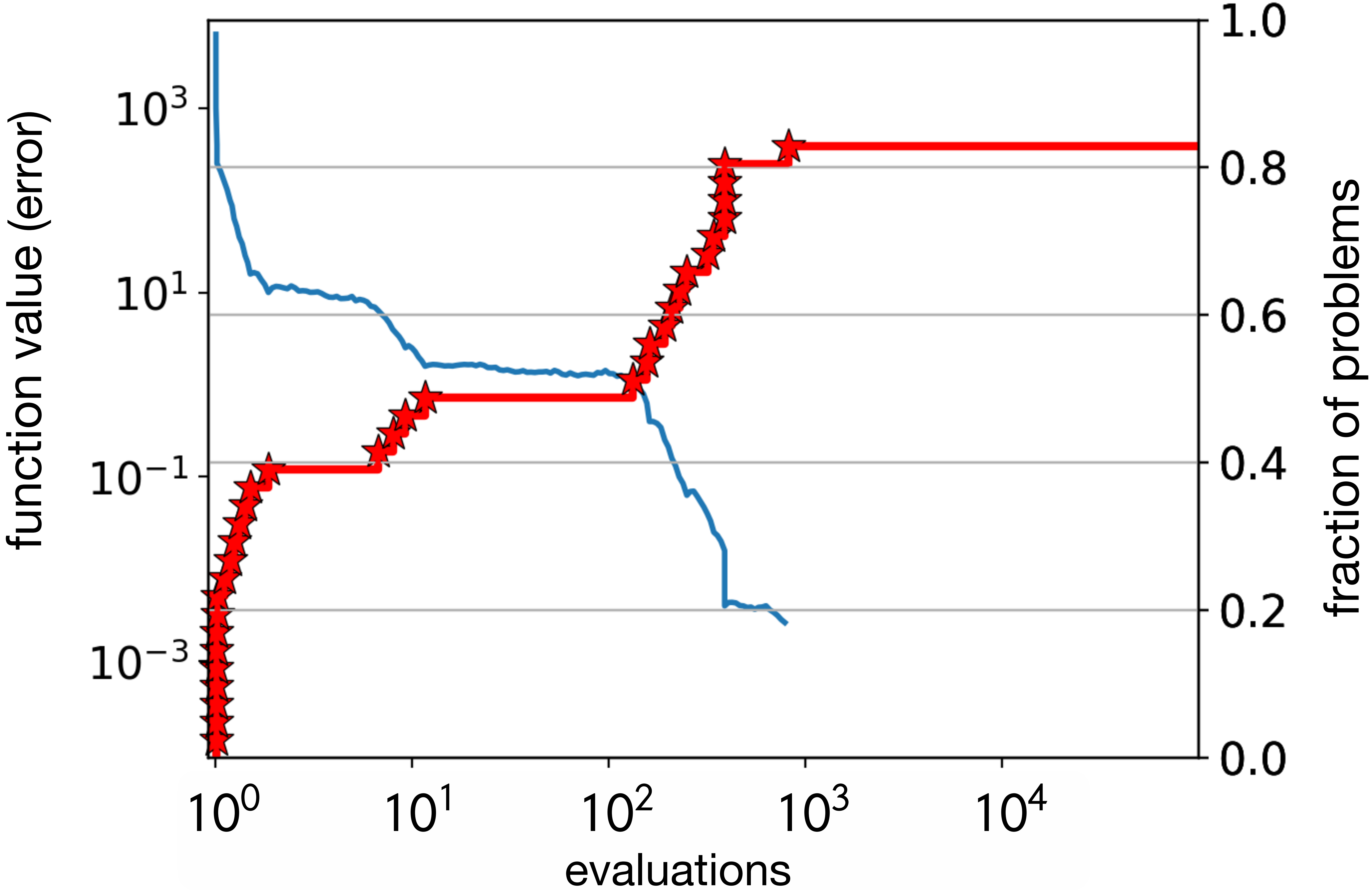
Hansen et al. 2022. Anytime performance assessment in blackbox optimization benchmarking. *IEEE Trans. on EC*, 26(6).

From a Convergence Graph to the Empirical Runtime Distribution



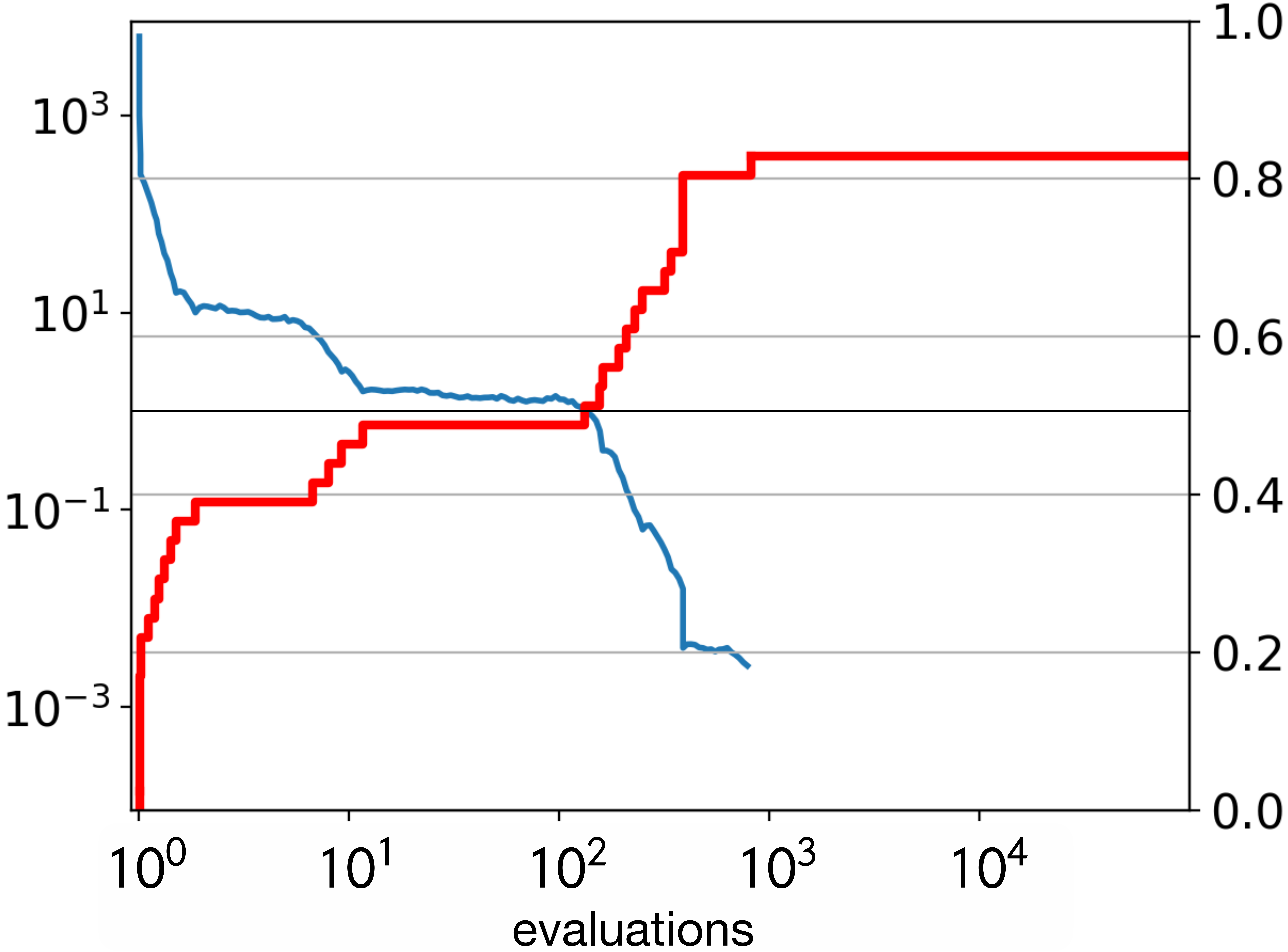
Hansen et al. 2022. Anytime performance assessment in blackbox optimization benchmarking. *IEEE Trans. on EC*, 26(6).

From a Convergence Graph to the Empirical Runtime Distribution



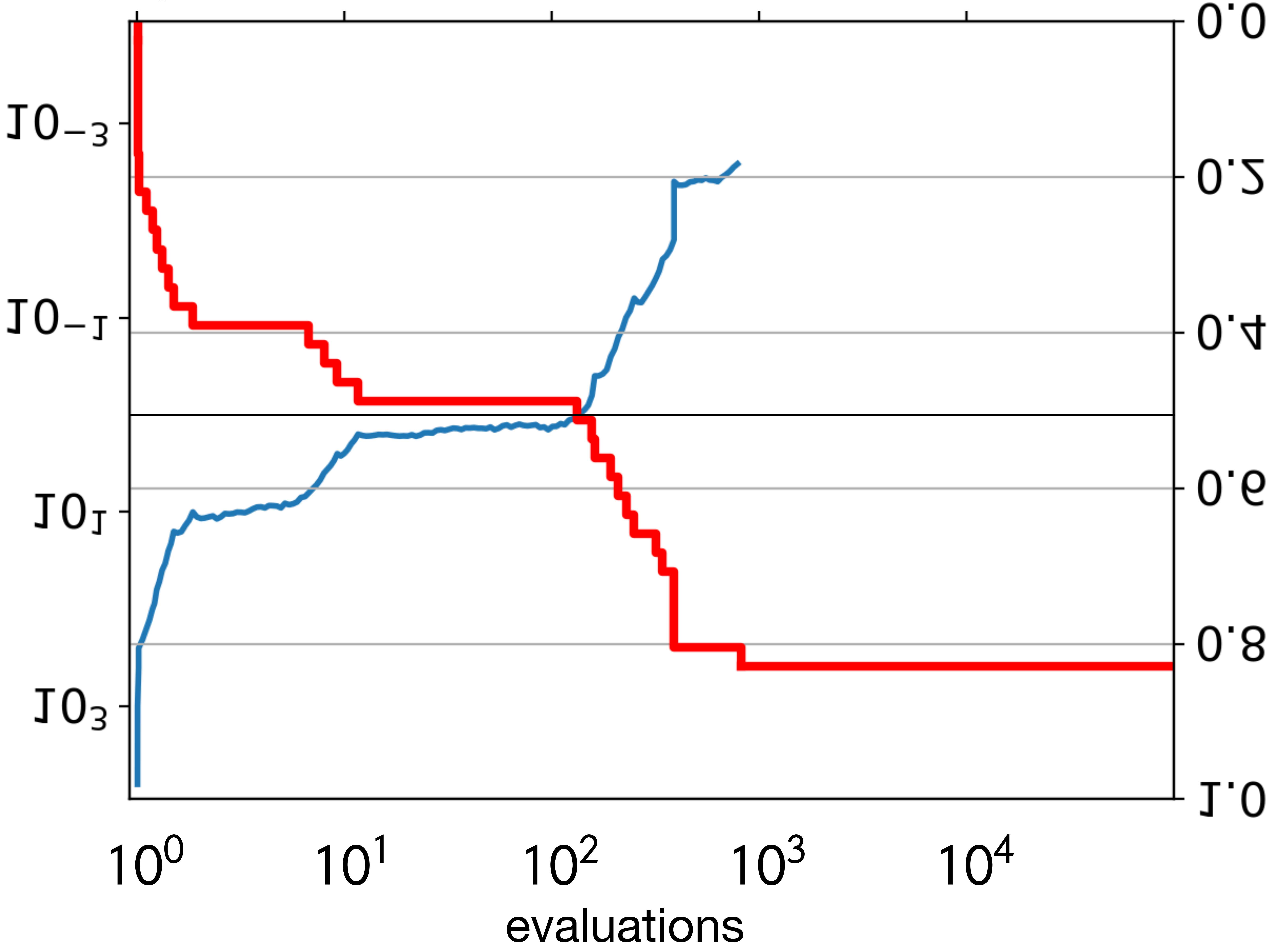
Hansen et al. 2022. Anytime performance assessment in blackbox optimization benchmarking. *IEEE Trans. on EC*, 26(6).

From a Convergence Graph to the Empirical Runtime Distribution



Hansen et al. 2022. Anytime performance assessment in blackbox optimization benchmarking. *IEEE Trans. on EC*, 26(6).

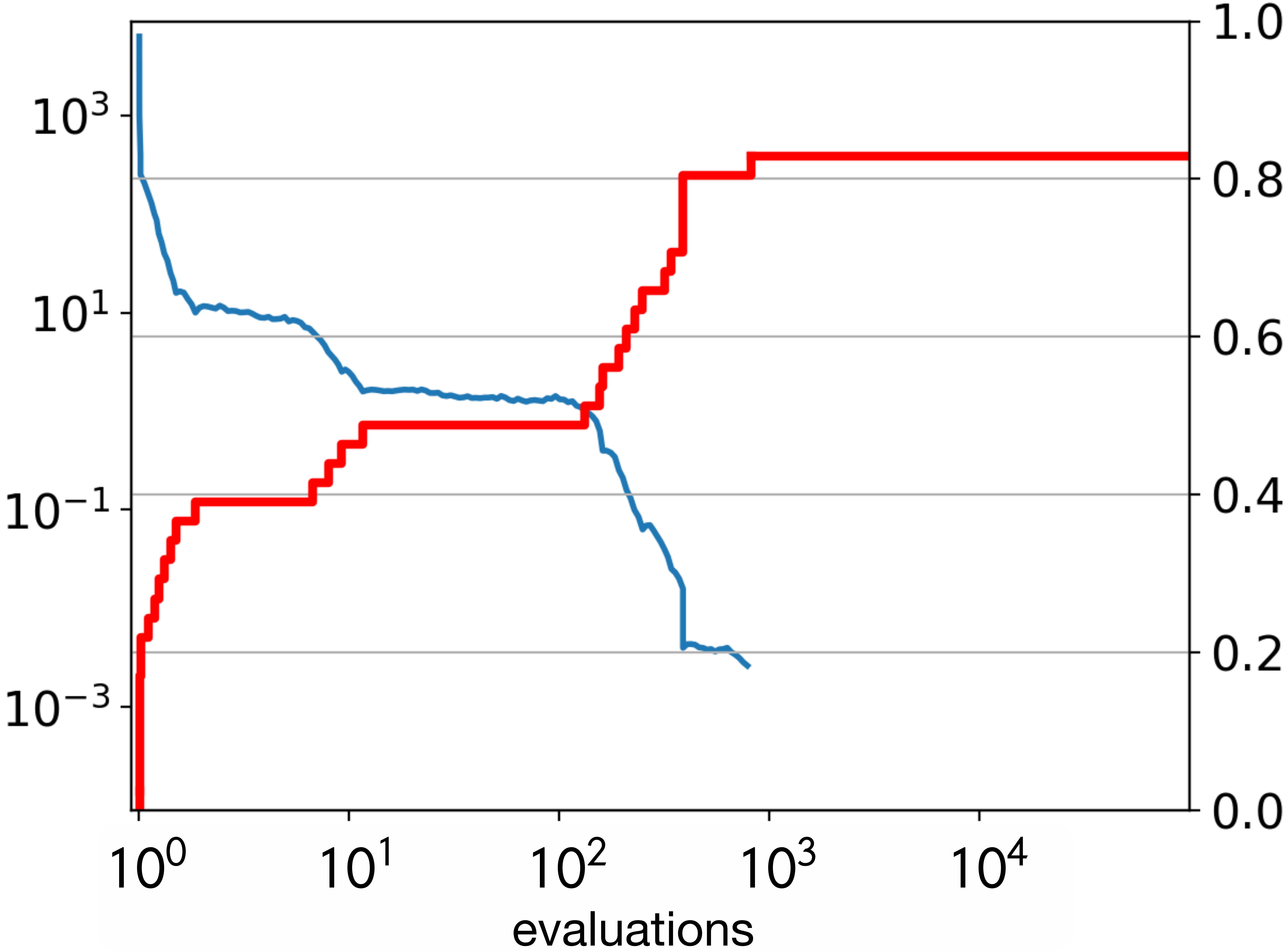
From a Convergence Graph to the Empirical Runtime Distribution



Hansen et al. 2022. Anytime performance assessment in blackbox optimization benchmarking. *IEEE Trans. on EC*, 26(6).

From a Convergence Graph to the Empirical Runtime Distribution

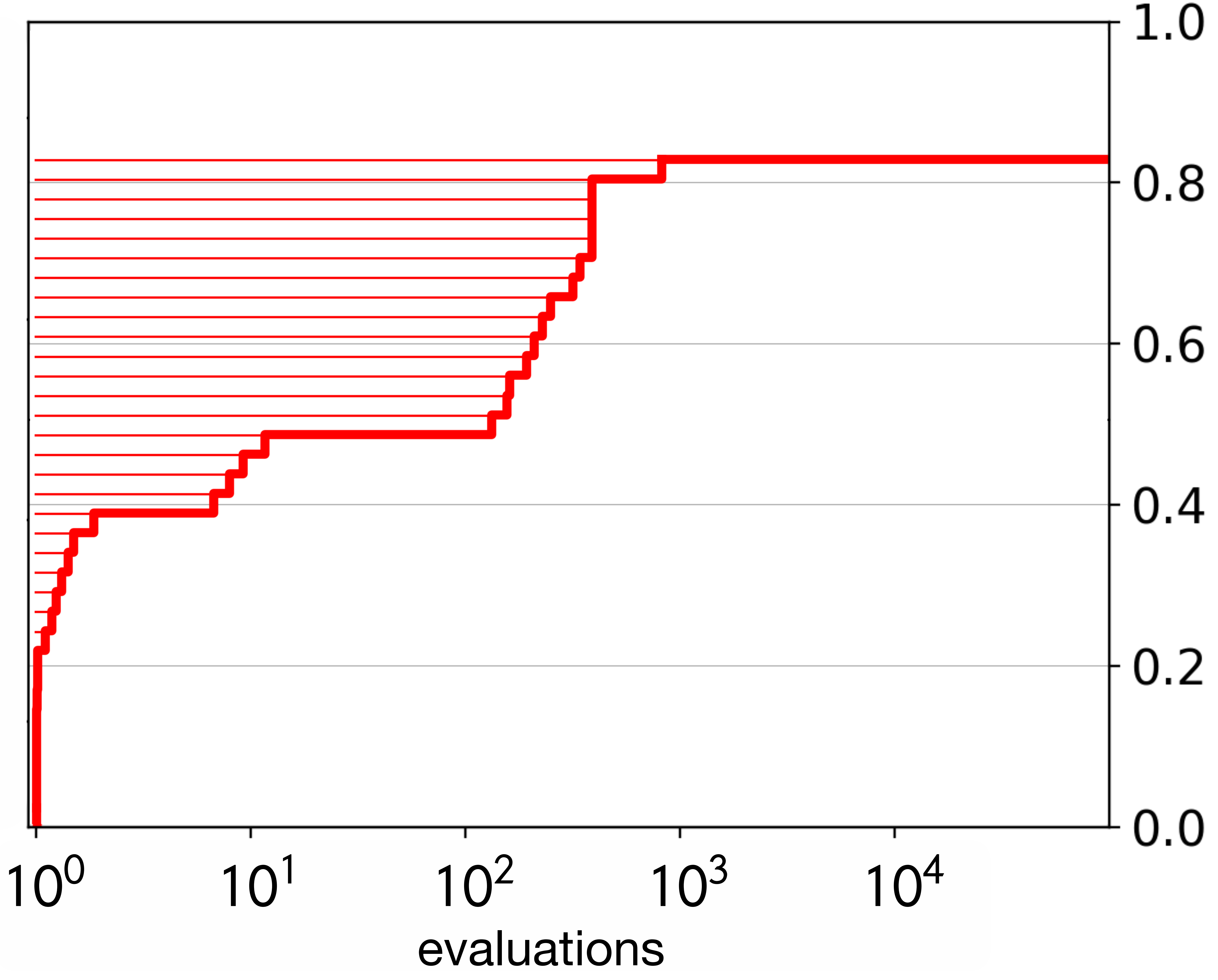
when we maximize (instead of minimize), the graph can be considered as an empirical runtime distribution *as is*



Hansen et al. 2022. Anytime performance assessment in blackbox optimization benchmarking. *IEEE Trans. on EC*, 26(6).

From a Convergence Graph to the Empirical Runtime Distribution

when we maximize (instead of minimize), the graph can be considered as an empirical runtime distribution *as is*

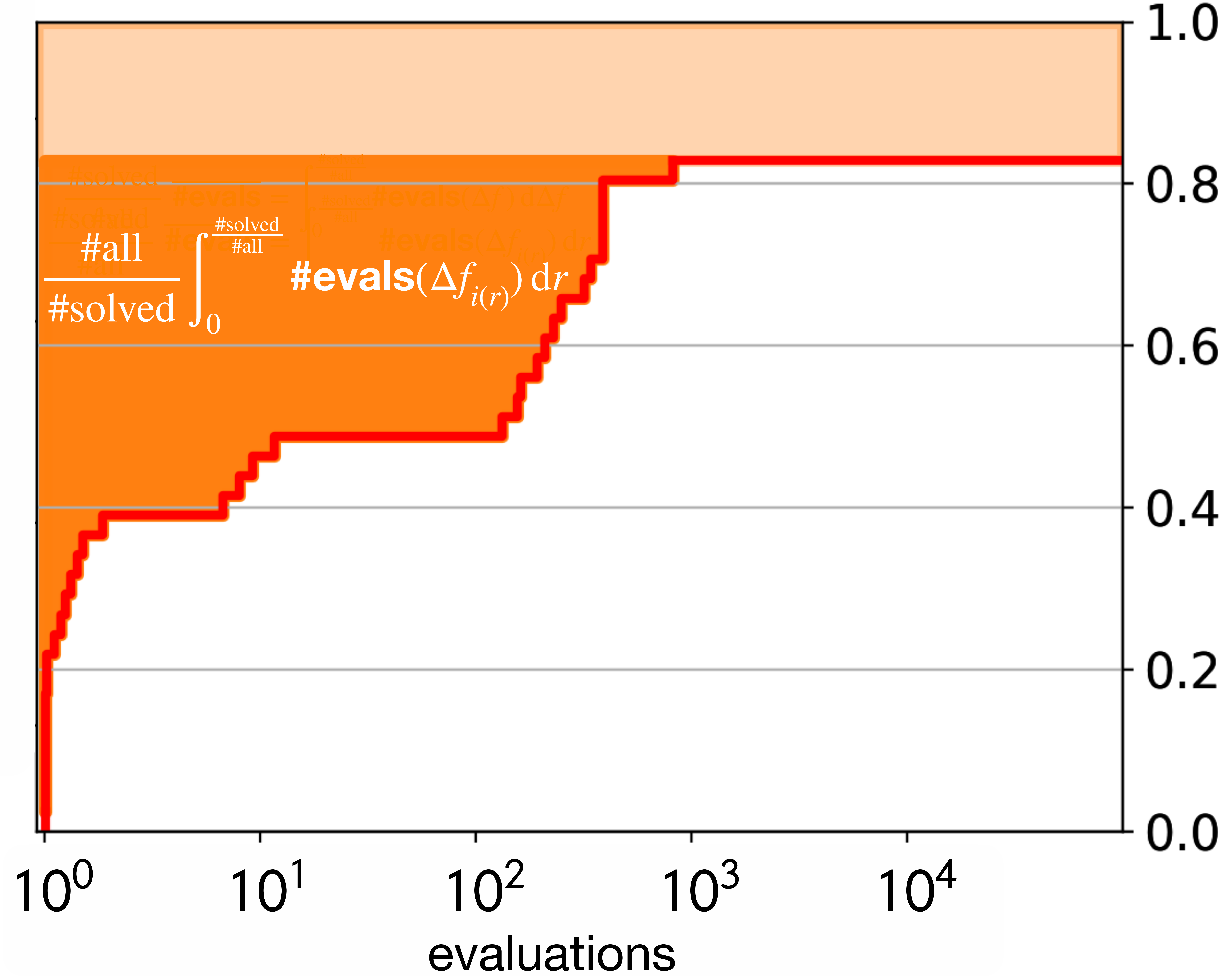


Hansen et al. 2022. Anytime performance assessment in blackbox optimization benchmarking. *IEEE Trans. on EC*, 26(6).

Empirical Runtime Distribution and area above the curve

the area above the curve represent a (truncated) average runtime

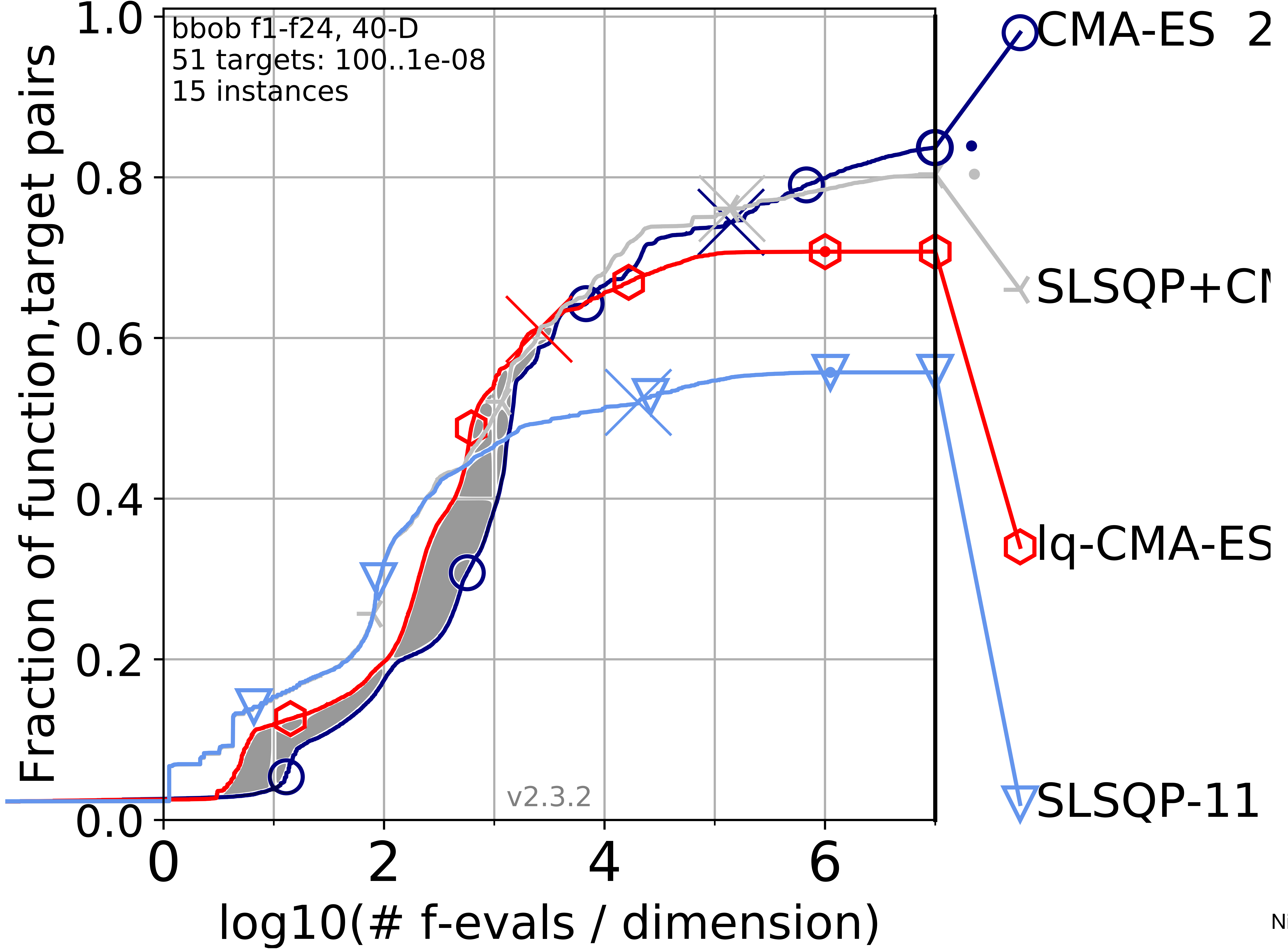
When the x-axis is in log-scale, the area is the (truncated) geometric average



Hansen et al. 2022. Anytime performance assessment in blackbox optimization benchmarking. *IEEE Trans. on EC*, 26(6).

Aggregated Runtime Distributions

Hansen 2019. A Global Surrogate Assisted CMA-ES. *GECCO '19*.



Nikolaus Hansen, Inria

Recall: the infamous p-value

Is the *probability for the observed data to be observed when the null hypothesis H_0 is true*

$$p = P(\text{observed data} \mid H_0)$$

we have $p \sim \mathcal{U}[0,1]$ when the data are sampled according to H_0

We are usually interested in rejecting H_0 with a small error, that is, we "desire"

$$P(H_0 \mid \text{observed data}) \ll 1$$

.

Common practice: we specify a threshold of "statistical significance", often 0.05, and reject H_0 when $p < 0.05$.

- Wasserstein et al. 2019: "*We conclude, based on our review of the articles in this special issue and the broader literature, that it is* ***time to stop using the term "statistically significant" entirely****. Nor should variants such as "significantly different," "p < 0.05," and "nonsignificant" survive, […] however, we are not recommending that the calculation and use of continuous p-values be discontinued. Where* ***p-values*** *are used, they* ***should be reported as continuous quantities*** *(e.g., p = 0.08).*"

Moving to a World Beyond "$p < 0.05$". *The American Statistician*, 73, S1.

- Amrhein et al. + 800 signatories, 2019: *"We agree, and call for the entire concept of* ***statistical significance to be abandoned****. […] we are calling for a stop to the use of P values* ***in the conventional, dichotomous way*** *— to decide whether a result refutes or supports a scientific hypothesis."*

Retire statistical significance. Scientists rise up against statistical significance. *Nature*, 567(7748).

- Cockburn et al. 2020: "*misuse of* ***statistical significance*** *as the standard of evidence for experimental success has been identified as a key contributor in the replication crisis."*

Threats of a replication crisis in empirical computer science. *Communications of the ACM*, 63(8).

A threshold of "statistical significance"

- creates a false dichotomy (significant vs not) $\Longrightarrow$ a mistaken mindset and mistaken conclusions

- fuels the replication crisis: passing (or not passing) the threshold leads to mistaken conclusions $\Longrightarrow$ replication fails

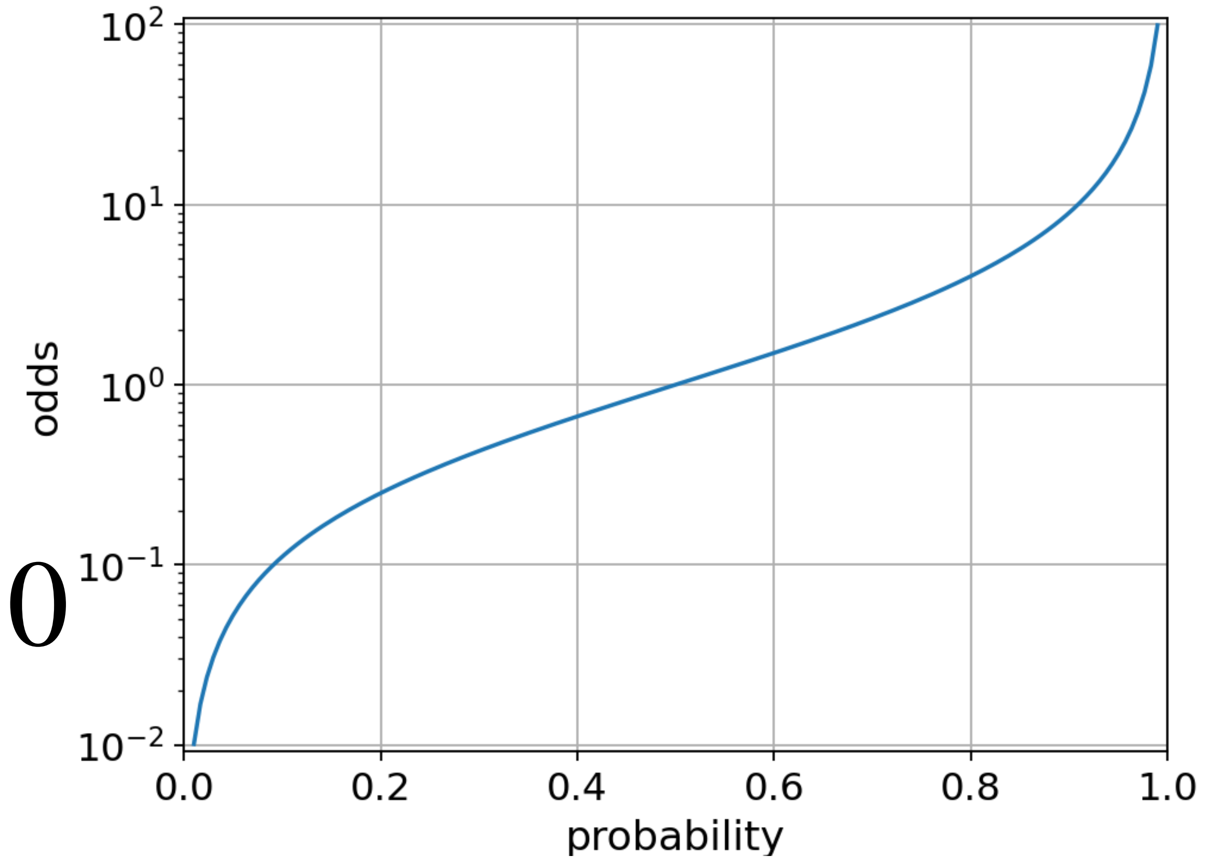
- any standard threshold value makes a silent (and oftentimes wrong) assumption on the prior probability of H_0

- adds no new information

  unless a case-specific argument is made for a case-specific value

How to use and *not* misuse the p-value?

Recall: The Odds

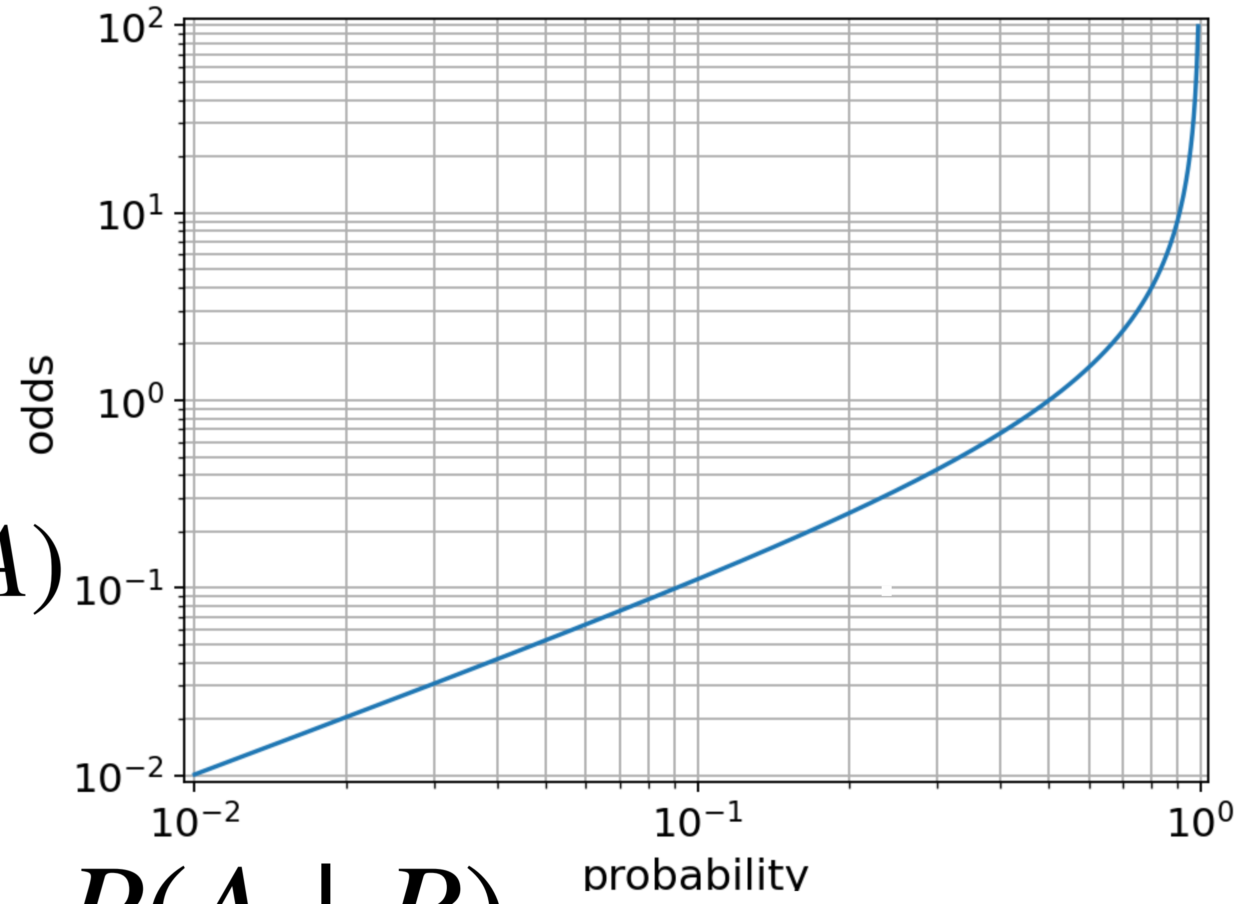


- The odds of A are defined as $o(A) = \dfrac{P(A)}{P(\neg A)} = \dfrac{P(A)}{1 - P(A)} \geq P(A) \geq 0$

- Probabilities and odds can be used interchangeably (whichever is more convenient)

  because there is a monotone bijection $p \in [0,1] \mapsto o(p) = \dfrac{p}{1-p} \in [0,\infty]$

  and $o \in [0,\infty] \mapsto p(o) = \dfrac{o}{o+1} \in [0,1]$



- For values close to zero, $o(A) \approx P(A)$ because the relative "error"

  $$\frac{|o(A) - P(A)|}{P(A)} = o(A)$$

  we have $o(A) = P(A) + o(A)P(A)$

- Correspondingly, the conditional or posterior odds are $o(A \mid B) = \dfrac{P(A \mid B)}{1 - P(A \mid B)}$

Nikolaus Hansen, Inria

From data to knowledge

$$o(H_0 \mid D) = \frac{P(H_0 \mid D)}{P(\neg H_0 \mid D)}, \quad o(H_0) = \frac{P(H_0)}{P(\neg H_0)}$$

The (famous) **Bayes' Rule in "odds form"** reads

$$\underbrace{o(H_0 \mid D)}_{\text{posterior odds}} = \underbrace{o(H_0)}_{\text{prior odds}} \times \underbrace{\frac{P(D \mid H_0)}{P(D \mid \neg H_0)}}_{\text{Bayes factor}}$$

Proof: Bayes' Theorem reads $P(H_0 \mid D) = P(H_0)\frac{P(D \mid H_0)}{P(D)}$ and likewise $P(\neg H_0 \mid D) = P(\neg H_0)\frac{P(D \mid \neg H_0)}{P(D)}$ and we divide the two equations.

Sources:
https://www.lesswrong.com/tag/odds
https://www.lesswrong.com/tag/log-odds
https://en.wikipedia.org/wiki/Bayes'_theorem#Bayes'_rule_in_odds_form

Nikolaus Hansen, Inria

From data to knowledge

$$o(H_0 \mid D) = \frac{P(H_0 \mid D)}{P(\neg H_0 \mid D)}, \quad o(H_0) = \frac{P(H_0)}{P(\neg H_0)}$$

The (famous) **Bayes' Rule in "odds form"** reads

$$\underbrace{o(H_0 \mid D)}_{\text{posterior odds}} = \underbrace{o(H_0)}_{\text{prior odds}} \times \underbrace{\frac{P(D \mid H_0)}{P(D \mid \neg H_0)}}_{\text{Bayes factor}}$$

The posterior odds are the odds to ***mistakenly*** **reject** H_0

which is close to the respective probability when $o(H_0 \mid D)$ is small

Nikolaus Hansen, Inria

From data to knowledge

$$o(H_0 \mid D) = \frac{P(H_0 \mid D)}{P(\neg H_0 \mid D)}, \quad o(H_0) = \frac{P(H_0)}{P(\neg H_0)}$$

The (famous) **Bayes' Rule in "odds form"** reads

$$\overset{10\,:\,1}{\underbrace{o(H_0 \mid D)}_{\text{posterior odds}}} = \overset{1000\,:\,1}{\underbrace{o(H_0)}_{\text{prior odds}}} \times \overset{\frac{1/100}{1}}{\underbrace{\frac{P(D \mid H_0)}{P(D \mid \neg H_0)}}_{\text{Bayes factor}}}$$

The posterior odds are the odds to ***mistakenly* reject** H_0

which is close to the respective probability when $o(H_0 \mid D)$ is small

From data to knowledge

$$o(H_0 \mid D) = \frac{P(H_0 \mid D)}{P(\neg H_0 \mid D)}, \quad o(H_0) = \frac{P(H_0)}{P(\neg H_0)}$$

Inserting the significance p-value in place of the nominator $P(D \mid H_0) \approx p$

$$\underbrace{o(H_0 \mid D)}_{\text{posterior odds}} = \underbrace{o(H_0)}_{\text{prior odds}} \times \underbrace{\frac{P(D \mid H_0)}{P(D \mid \neg H_0)}}_{\text{Bayes factor}}$$

Nikolaus Hansen, Inria

From data to knowledge

$$o(H_0 \mid D) = \frac{P(H_0 \mid D)}{P(\neg H_0 \mid D)}, \quad o(H_0) = \frac{P(H_0)}{P(\neg H_0)}$$

Inserting the significance p-value in place of the nominator $P(D \mid H_0) \approx p$

$$\underbrace{o(H_0 \mid D)}_{\text{posterior odds}} \approx \underbrace{o(H_0)}_{\text{prior odds}} \times \underbrace{\frac{p}{P(D \mid \neg H_0)}}_{\text{Bayes factor}}$$

and assuming the denominator $P(D \mid \neg H_0) \approx 1/2$ (D is a typical observation when $\neg H_0$ is true)

From data to knowledge

$$o(H_0 \mid D) = \frac{P(H_0 \mid D)}{P(\neg H_0 \mid D)}, \quad o(H_0) = \frac{P(H_0)}{P(\neg H_0)}$$

yields

$$\underbrace{o(H_0 \mid D)}_{\text{posterior odds}} \quad \approx \quad \underbrace{o(H_0)}_{\text{prior odds}} \times 2p$$

as a rough approximation for the posterior odds of H_0.

This is how to use the p-value — as the *amount of evidence* with which we can *update* our confidence (or lack thereof) in H_0.

Nikolaus Hansen, Inria

From data to knowledge

$$o(H_0 \mid D) = \frac{P(H_0 \mid D)}{P(\neg H_0 \mid D)}, \quad o(H_0) = \frac{P(H_0)}{P(\neg H_0)}$$

yields

$$\underbrace{o(H_0 \mid D)}_{\text{posterior odds}} \approx \underbrace{o(H_0)}_{\text{prior odds}} \times 2p$$

as a rough approximation for the posterior odds of H_0.

Sagan 1979: Extraordinary claims require extraordinary evidence.

Sagan 1979. Broca's Brain.

The claim to reject H_0 when the Odds$(H_0) \gg 1$ is extraordinary and **requires** $p \ll$ **Odds**$(H_0)^{-1}$.

Independent repetition/replication

$$o(H_0 \mid D) = \frac{P(H_0 \mid D)}{P(\neg H_0 \mid D)}, \quad o(H_0) = \frac{P(H_0)}{P(\neg H_0)}$$

This works for independent replications too!

$$\underbrace{o\left(H_0 \mid \bigcap_{i=1}^{k} D_i\right)}_{\text{posterior odds}} \approx \underbrace{o(H_0)}_{\text{prior odds}} \times \underbrace{\prod_{i=1}^{k} \frac{p_i}{P(D_i \mid \neg H_0)}}_{\text{Bayes factors}}$$

$$\approx \underbrace{o(H_0)}_{\text{prior odds}} \times \prod_{i=1}^{k} 2p_i$$

The probability of H_0 vanishes geometrically fast with the number of replications.

Nikolaus Hansen, Inria

From data to knowledge

The observed p-value indicates by how much we should ***update*** our confidence in H_0 (not: how confident we should be in H_0)

$$\underbrace{\text{Odds}(H_0 \mid D)}_{\text{posterior odds}} \approx \underbrace{\text{Odds}(H_0)}_{\text{prior odds}} \times 2p$$

If we do not provide an estimate for the prior odds, we have no argument to reject H_0 (that's perfectly fine)

a small p stands on its own merits: we *can* conclude that the odds for H_0 have decreased by a factor of about $2p$

If we improved a well-established state-of-the-art algorithm or invented cold fusion or found a room-temperature superconductor at atmospheric pressures, the prior odds of H_0 are usually high, say, e.g. 10^4.

the higher the prior odds for H_0, the more exceptional or surprising is the scientific result

to accept $\neg H_0$ with the same confidence we before had in H_0, we need $p \approx P(\neg H_0)^2$

Recommendations: Quantify…

- Always quantify effect size (if at all possible).

  Specifically, don’t rank algorithms (don’t say “A was the fastest”, say “A was 3% faster than the second fastest” or “A and B essentially performed the same”).

  - Don’t ask yourself: ~~Was deep learning better than random forests (on this application)?~~

  - Ask instead: **How much better** was deep learning compared to random forests (on this application)?

Recommendations: Quantify…

- Always quantify effect size (if at all possible).
- Don't write (ever again) "statistically significant", instead report p-values (as a quantification of evidence).
  - Don't ask yourself: ~~Was the difference statistically significant?~~
  - Ask instead: **How small** was the p-value (approximately)?
  - Remember: our confidence in H_0 *change* **by a factor** of about $2p$ (decrease when $p < 1/2$)

not: ~~$p < 0.05 \implies$ odds to reject H_0 *mistakenly* are small~~

Nikolaus Hansen, Inria

Recommendations: Quantify…

- Always quantify effect size (if at all possible).
- Don't write (ever again) "statistically significant", instead report p-values (as a quantification of evidence).
- Wait for replications before to conclude that a result is replicable (science is incremental)

  - Don't ask yourself: ~~Is this paper replicable?~~
  - Ask instead: How often has this result **been replicated**? What is the current (posterior) odds for H_0? $\Longrightarrow$ quantification of the confidence in replicability.