From Data to Knowledge

Nikolaus Hansen
Inria & École polytechnique, France

September 2026

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 form a scientific claim, science aims to make universal claims

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

$\Longrightarrow$ replicability becomes the criterion of demarcation

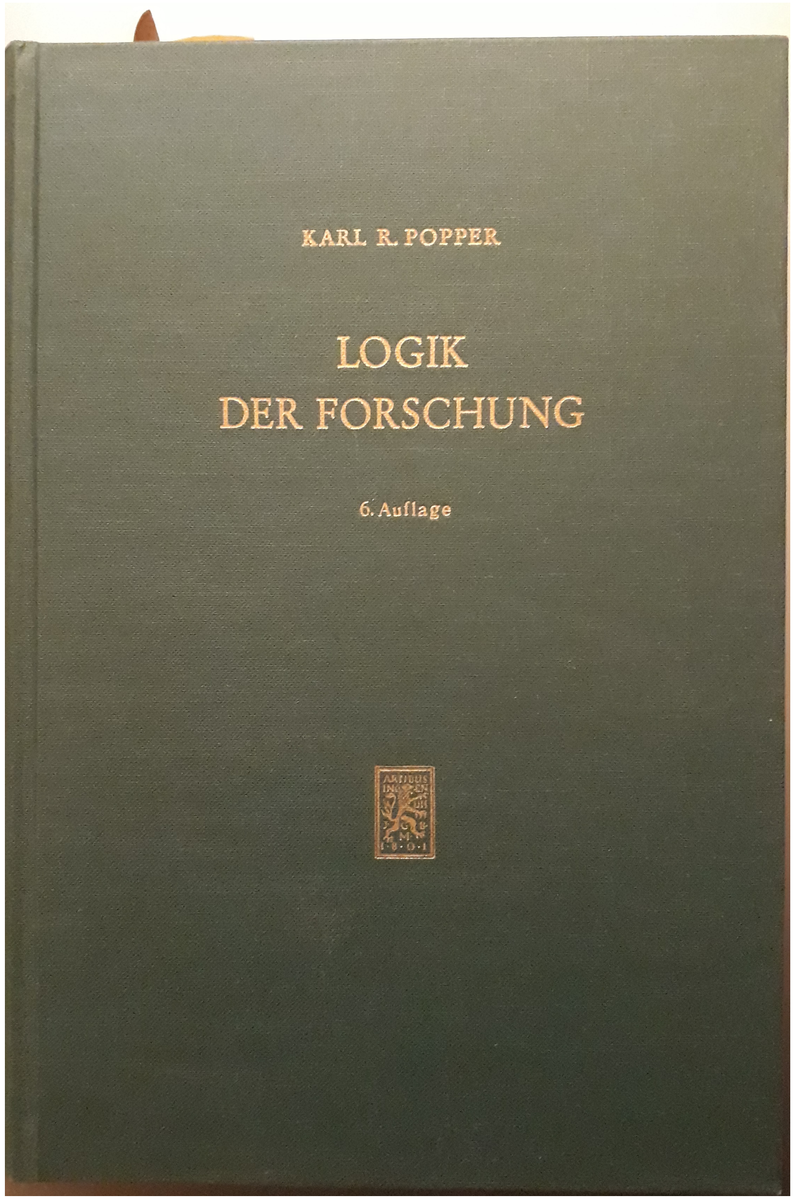
$\Longrightarrow$ claims are universal and falsifiable (and supported by failed falsifications)

$\Longrightarrow$ 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

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 form a scientific claim, science aims to make universal claims

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

$\Longrightarrow$ replicability becomes the criterion of demarcation

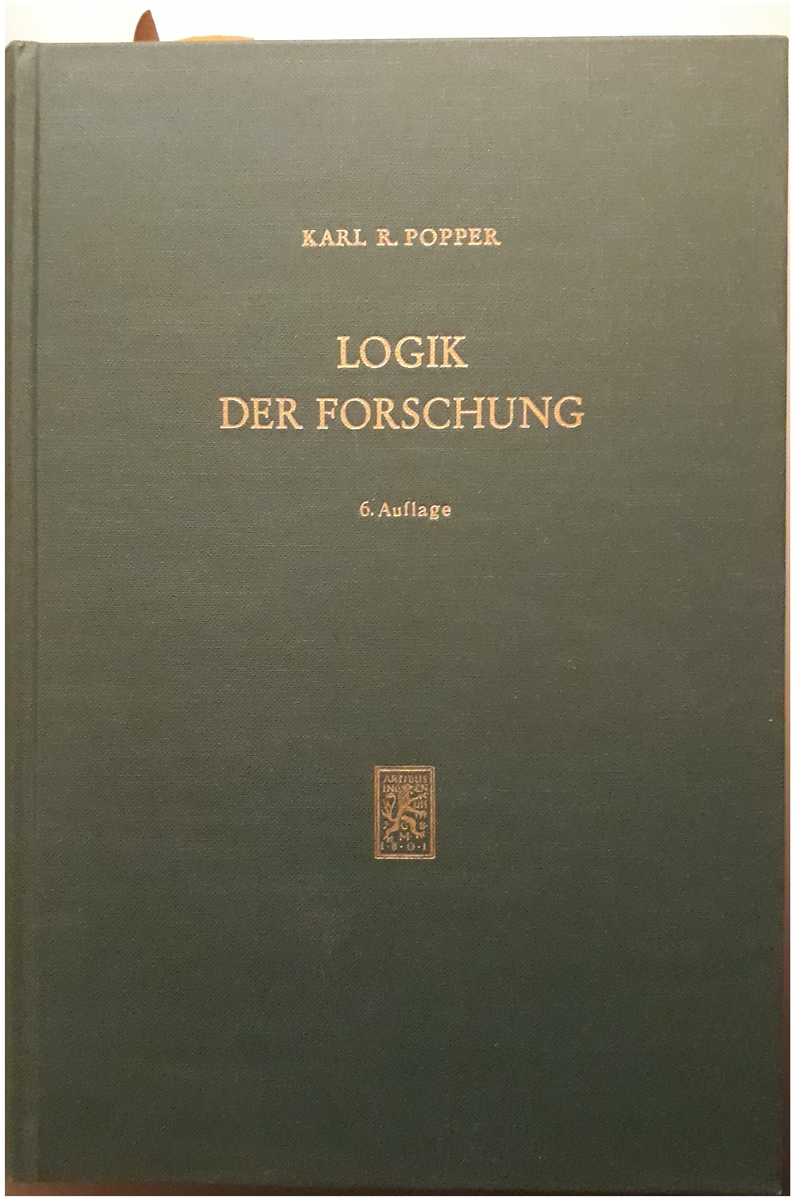
$\Longrightarrow$ claims are universal and falsifiable (and supported by failed falsifications)

$\Longrightarrow$ 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

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 form a scientific claim, science aims to make universal claims

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

$\Longrightarrow$ replicability becomes the criterion of demarcation

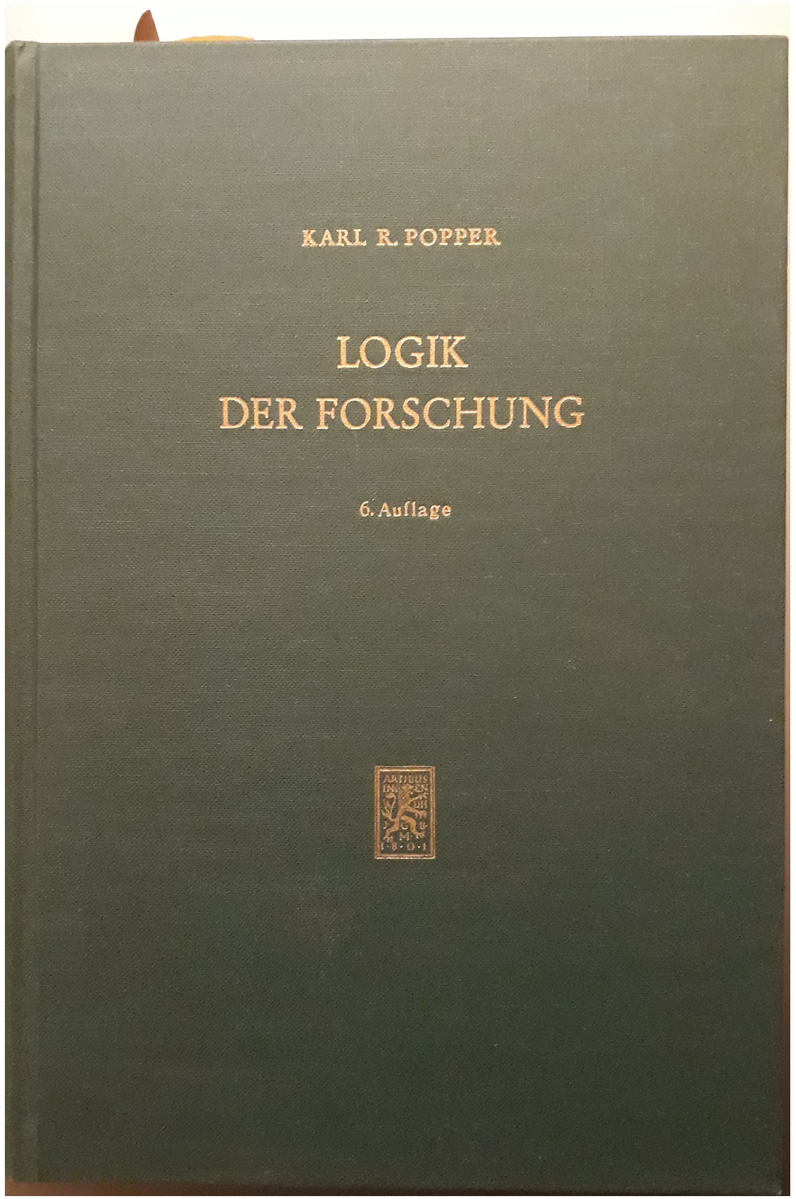
$\Longrightarrow$ claims are universal and falsifiable (and supported by failed falsifications)

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

not only standing on the shoulders of giants

$\Longrightarrow$ consensus (“undeniable” wisdom)

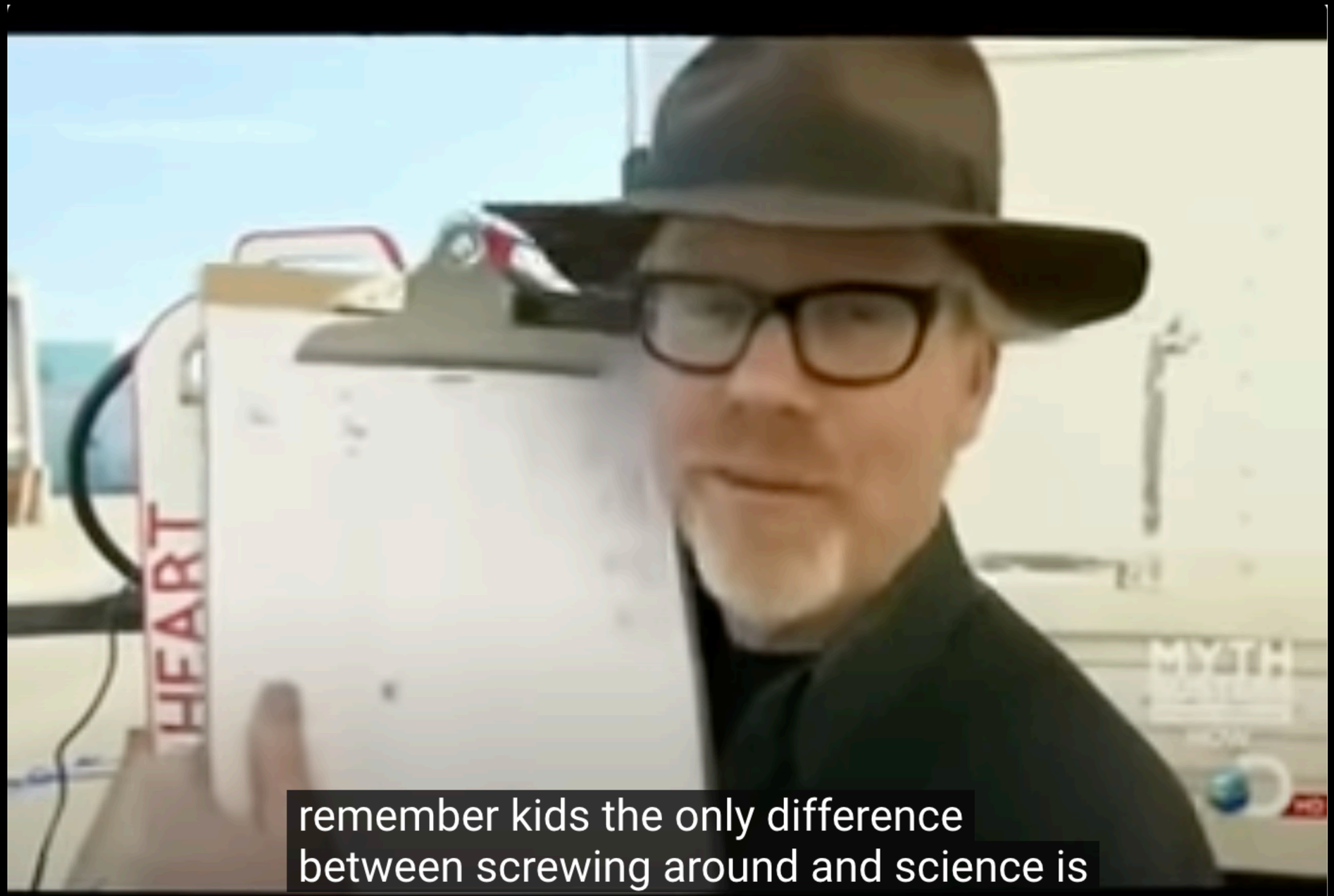
$\Longrightarrow$ knowledge



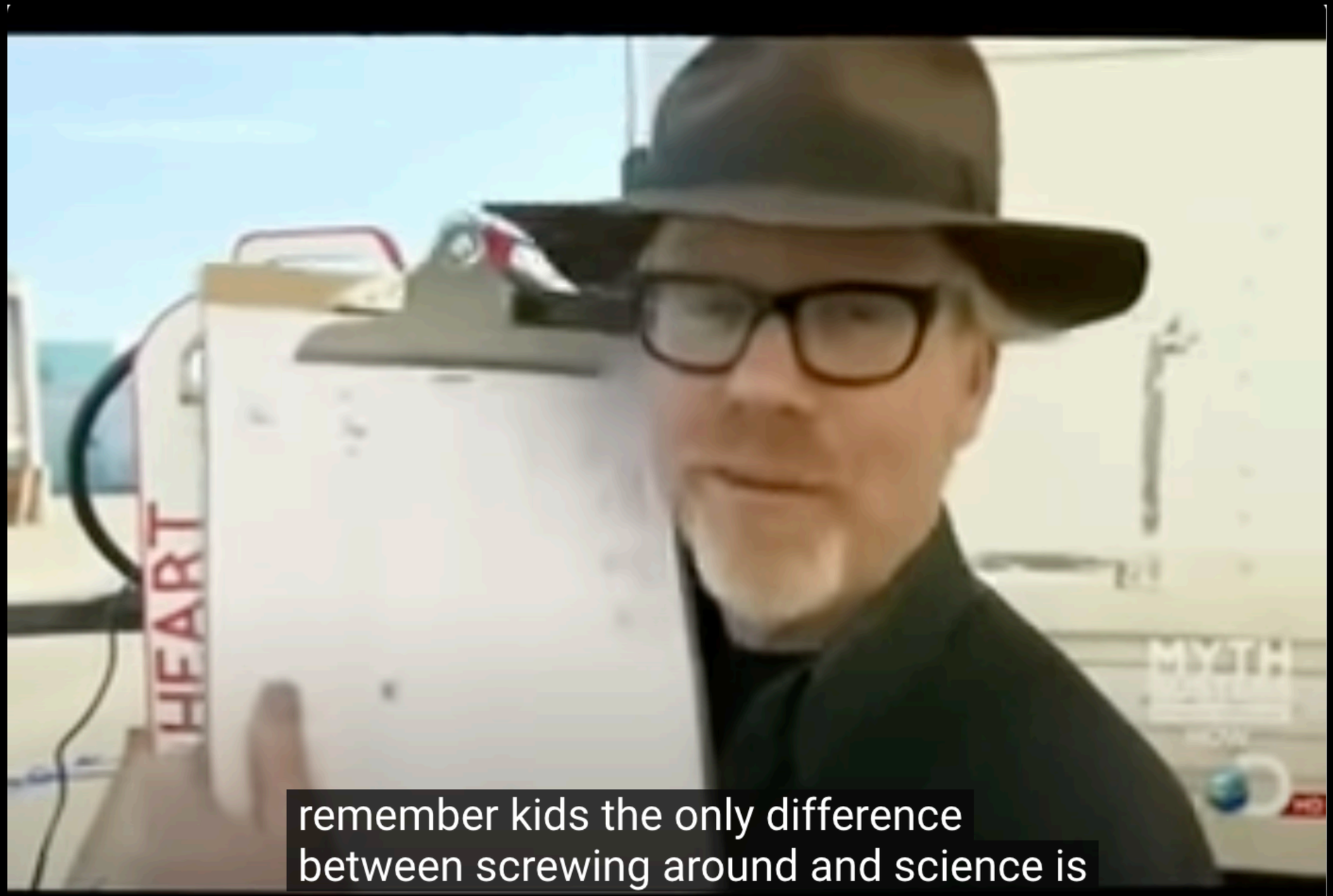
Popper 1959. The logic of scientific discovery

https://youtu.be/b240PGCMwV0

The Key to Science



— Adam Savage (on YouTube)



— Adam Savage (on YouTube)

writing it down

The Import Distinction

- **Generating (promising) hypotheses** (e.g. finding a "good" algorithm) does *not require any* methodological or scientific standards!

  We are permitted to do as quick and dirty experiments as we see fit!
  We are permitted to use or ignore any information or data!
  Quick and dirty "testing" is often part of the hypothesis generation process.

- **Testing hypotheses** = *evaluating the truth value of a candidate hypothesis* (e.g. assessing the performance of an algorithm) requires rigorous methodological standards *and* documentation to create credible evidence.

  Specifically, we are not permitted to use any data used to *generate* the hypothesis.

  Any testing only ever *changes* (updates) our current estimate of the truth value!

  There is never a "proof", there is only overwhelming evidence.

We usually spent much more time in trying to generate *promising* hypotheses than in testing them, hence "outside" of any concise prescriptions of the scientific method.

Formulating a groundbreaking algorithm is much more difficult than assessing whether the performance of an algorithm *is* groundbreaking.

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).

General reasons (sorted by the impact of the “human factor”):

- fraud (rare)
- misrepresentation or exaggeration (common)
- mistakes (not rare)
- weak or flawed methodology (common)

  what must be considered “flawed” is a moving notion
- intrinsic to the process (not rare)

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).

General reasons even in the "ideal" world:

- mistakes (not rare)
- intrinsic to the process (not rare)

Implication: we need to find ways to deal with **the lack of** replicability no matter what.

Observing lack of replicability is not (always) a failure, it is (also and) in particular part of the success story of science.

Arguably, the negative connotation of “lack of replicability” is a misplaced connotation in the first place.

specifically…

Why are papers not replicable, specifically?

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 convenient for comparison/competitor algorithms

- selective/distorted/biased/misleading/exaggerating reporting

- fraud

The less obvious:

- small effect size

  $\Longrightarrow$ false negative or misleading positive outcome

- misinterpreting the p-value:

The less obvious:

- small effect size

$\Longrightarrow$ false negative or misleading positive outcome

- misinterpreting the p-value:
  - high "multiplicity"
    - large number of tested relationships (*p*-fishing)
    - many teams working independently in parallel
    - great flexibility in study design and analytical modes (methods)

    leads to *selection and publication bias* (by rarely reporting failures)
  - small prior ratio of true to false hypotheses (Odds$(H_0) \gg 1$, a good algorithm is difficult to improve)
  - confounding hypothesis *generating* data (HARKing) with hypothesis *testing* data (evidence)

these concern authors *and readers*

The less obvious:

- small effect size

$\Longrightarrow$ false negative or misleading positive outcome

- misinterpreting the p-value:
  - high "multiplicity"
    - large number of tested relationships (*p*-fishing)
    - many teams working independently in parallel
    - great flexibility in study design and analytical modes (methods)

    leads to *selection and publication bias* (by rarely reporting failures)
  - small prior ratio of true to false hypotheses (Odds$(H_0) \gg 1$, a good algorithm is difficult to improve)
  - confounding hypothesis *generating* data (HARKing) with hypothesis *testing* data (evidence)

these concern authors *and readers*

- misinterpreting the p-value:
  - high "multiplicity"
    - large number of tested relationships (*p*-fishing)
    - many teams working independently in parallel
    - great flexibility in study design and analytical modes (methods)

    leads to *selection and publication bias* (by rarely reporting failures)
  - small prior ratio of true to false hypotheses (Odds$(H_0) \gg 1$, a good algorithm is difficult to improve)
  - confounding hypothesis *generating* data (HARKing) with hypothesis *testing* data (evidence)

these concern 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 "four horsemen" of irreproducibility

Bishop 2019. Rein in the four horsemen of irreproducibility. *Nature*, 568(7753).

The “four horsemen”:

- selection and publication bias (negative results are rarely published)
- low statistical power (small number of samples)
- p-value hacking (repeated statistical analyses)
- HARKing (generating Hypotheses After the Results are Known)

  and claiming falsely that the results *confirm* the hypothesis

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.*"

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 information.

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 information.

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 information.

What can we do, specifically?

Besides the obvious (that is, with the above in mind, writing better papers):

- Individually:
  - **read** papers properly

    a single paper can (almost) never be considered as the ground truth
    scrutinize the methodology

    estimate the prior and posterior odds of H_0
  - require a second, independent source

    a common standard in journalism
- Collectively: allow for
  - publication of **replication studies**

    explicitly distinguish between *hypotheses developing* and *hypotheses testing* procedures [Gigerenzer 2018]
  - **preregistration** of new studies

    that is, publication *by methodology and relevance* rather than by results or p-values

Quantification

quantify quantify quantify

Quantification

Quantify.

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

Sagans nine tools of baloney detection are

1.There must be independent confirmation of the "facts" wherever possible.
2. Encourage substantive debate on the evidence by knowledgeable proponents of all points of view.
3. Arguments from authority carry little weight.
4. Spin more than one hypothesis.
5. Try not to get overly attached to a hypothesis just because it's yours (form of confirmation bias)
6. **Quantify**.
7. If there is a chain of argument, *every* link in the chain must work, including the premise.
8. Occam's Razor.
9. Always ask whether the hypothesis can be, at least in principle, falsified.

Quantification of knowledge certainty

Quantify.

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

Sagans nine tools of baloney detection are

1.There must be independent confirmation of the "facts" wherever possible.
2. Encourage substantive debate on the evidence by knowledgeable proponents of all points of view.
3. Arguments from authority carry little weight.
4. Spin more than one hypothesis.
5. Try not to get overly attached to a hypothesis just because it's yours (form of confirmation bias)
6. **Quantify**.
7. If there is a chain of argument, *every* link in the chain must work, including the premise.
8. Occam's Razor.
9. Always ask whether the hypothesis can be, at least in principle, falsified.

Bayes' theorem in odds form

$$\frac{P(H_0 \mid D)}{P(H_1 \mid D)} = \frac{P(H_0)}{P(H_1)} \times \frac{P(D \mid H_0)}{P(D \mid H_1)}$$

H: hypothesis
D: data
P: probability

Everything else is commentary

H: hypothesis
D: data
P: probability

Sources:
https://www.lesswrong.com/tag/log-odds
https://en.wikipedia.org/wiki/Bayes'_theorem#In_odds_form
Ly et al 2016 https://doi.org/10.1016/j.jmp.2015.06.004

Bayes' theorem in odds form

$$\frac{P(H_0 \mid D)}{P(H_1 \mid D)} = \frac{P(H_0)}{P(H_1)} \times \frac{P(D \mid H_0)}{P(D \mid H_1)}$$

H: hypothesis
D: data
P: probability

Everything else is commentary

H: hypothesis
D: data
P: probability

Sources:
https://www.lesswrong.com/tag/log-odds
https://en.wikipedia.org/wiki/Bayes'_theorem#In_odds_form
Ly et al 2016 https://doi.org/10.1016/j.jmp.2015.06.004

Bayes' theorem in odds form

$$\frac{P(H_0 \mid D)}{P(H_1 \mid D)} = \frac{P(H_0)}{P(H_1)} \times \frac{P(D \mid H_0)}{P(D \mid H_1)}$$

H: hypothesis
D: data
P: probability

Everything else is commentary

H: hypothesis
D: data
P: probability

Sources:
https://www.lesswrong.com/tag/log-odds
https://en.wikipedia.org/wiki/Bayes'_theorem#In_odds_form
Ly et al 2016 https://doi.org/10.1016/j.jmp.2015.06.004

Bayes' theorem in odds form

$$\frac{P(H_0 \mid D)}{P(H_1 \mid D)} = \frac{P(H_0)}{P(H_1)} \times \frac{P(D \mid H_0)}{P(D \mid H_1)}$$

H: hypothesis
D: data
P: probability

Everything else is commentary

H: hypothesis
D: data
P: probability

Sources:
https://www.lesswrong.com/tag/log-odds
https://en.wikipedia.org/wiki/Bayes'_theorem#In_odds_form
Ly et al 2016 https://doi.org/10.1016/j.jmp.2015.06.004

prior "odds" of hypothesis H_0 versus H_1

$$\frac{P(H_0 \mid D)}{P(H_1 \mid D)} = \frac{P(H_0)}{P(H_1)} \times \frac{P(D \mid H_0)}{P(D \mid H_1)}$$

Everything else is commentary

H: hypothesis
D: data
P: probability

prior "odds" of hypothesis H_0 versus H_1

posterior "odds" of hypothesis H_0 versus H_1
given the data D

$$\underbrace{\frac{P(H_0 \mid D)}{P(H_1 \mid D)}}_{\text{posterior "odds"}} = \frac{P(H_0)}{P(H_1)} \times \frac{P(D \mid H_0)}{P(D \mid H_1)}$$

Everything else is commentary

H: hypothesis
D: data
P: probability

prior "odds" of hypothesis H_0 versus H_1

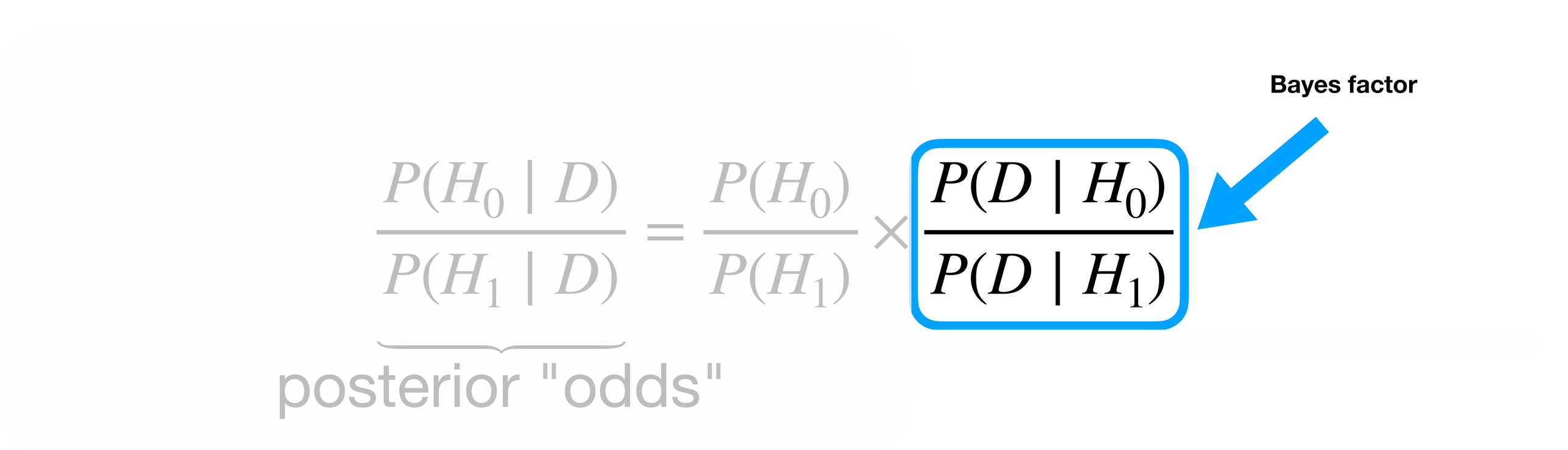
posterior "odds" of hypothesis H_0 versus H_1
given the data D

Bayes factor

$$\underbrace{\frac{P(H_0 \mid D)}{P(H_1 \mid D)}}_{\text{posterior "odds"}} = \frac{P(H_0)}{P(H_1)} \times \frac{P(D \mid H_0)}{P(D \mid H_1)}$$

Everything else is commentary

H: hypothesis
D: data
P: probability



Everything else is commentary

H: hypothesis
D: data
P: probability

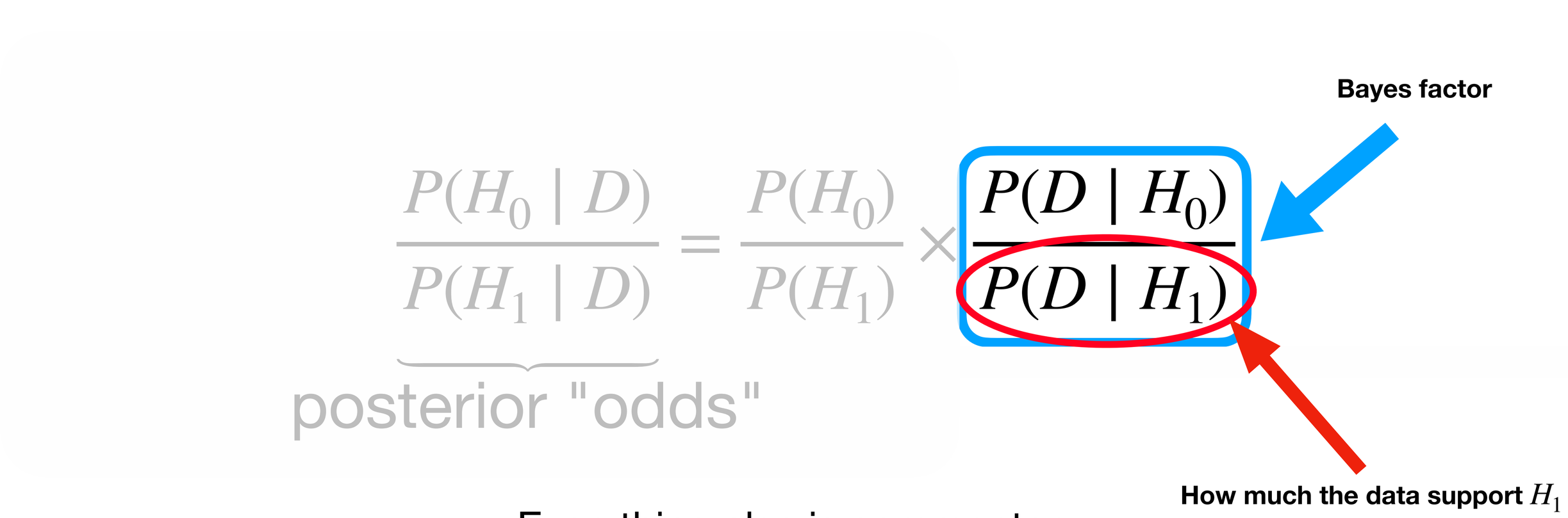
$$\underbrace{\frac{P(H_0 \mid D)}{P(H_1 \mid D)}}_{\text{posterior "odds"}} = \frac{P(H_0)}{P(H_1)} \times \frac{P(D \mid H_0)}{P(D \mid H_1)}$$

Bayes factor

How much the data support H_0

Everything else is commentary

H: hypothesis
D: data
P: probability



Everything else is commentary

H: hypothesis
D: data
P: probability

prior "odds" of hypothesis H_0 versus H_1

posterior "odds" of hypothesis H_0 versus H_1
given the data D

Bayes factor

$$\underbrace{\frac{P(H_0 \mid D)}{P(H_1 \mid D)}}_{\text{posterior "odds"}} = \frac{P(H_0)}{P(H_1)} \times \frac{P(D \mid H_0)}{P(D \mid H_1)}$$

Everything else is commentary

H: hypothesis
D: data
P: probability

prior "odds" of hypothesis H_0 versus H_1

posterior "odds" of hypothesis H_0 versus H_1
given the data D

Bayes factor

$$\underbrace{\frac{P(H_0 \mid D)}{P(H_1 \mid D)}}_{\text{posterior "odds"}} = \frac{P(H_0)}{P(H_1)} \times \frac{P(D \mid H_0)}{P(D \mid H_1)}$$

Special cases:

- $H_1 = \neg H_0$, then $\frac{P(H_0)}{P(H_1)} = \frac{P(H_0)}{P(\neg H_0)} = \frac{P(H_0)}{1 - P(H_0)}$ is the odds of H_0
- $H_1 = H_0 \cup \neg H_0$ recovers the classical formulation of the Bayesian theorem as $P(H_1 \mid D) = P(H_1) = 1$ and $P(D \mid H_1) = P(D)$

H: hypothesis
D: data
P: probability

prior "odds" of hypothesis H_0 versus H_1

posterior "odds" of hypothesis H_0 versus H_1
given the data D

Bayes factor

$$\underbrace{\frac{P(H_0 \mid D)}{P(H_1 \mid D)}}_{\text{posterior "odds"}} = \frac{P(H_0)}{P(H_1)} \times \frac{P(D \mid H_0)}{P(D \mid H_1)}$$

Special cases:

- $H_1 = \neg H_0$, then $\frac{P(H_0)}{P(H_1)} = \frac{P(H_0)}{P(\neg H_0)} = \frac{P(H_0)}{1 - P(H_0)}$ is the odds of H_0
- $H_1 = H_0 \cup \neg H_0$ recovers the classical formulation of the Bayesian theorem as $P(H_1 \mid D) = P(H_1) = 1$ and $P(D \mid H_1) = P(D)$

H: hypothesis
D: data
P: probability

prior "odds" of hypothesis H_0

posterior "odds" of hypothesis H_0 *given the data* D

Bayes factor

$$\underbrace{\frac{P(H_0 \mid D)}{P(\neg H_0 \mid D)}}_{\text{posterior "odds"}} = \frac{P(H_0)}{P(\neg H_0)} \times \frac{P(D \mid H_0)}{P(D \mid \neg H_0)}$$

Special cases:

- $H_1 = \neg H_0$, then $\frac{P(H_0)}{P(H_1)} = \frac{P(H_0)}{P(\neg H_0)} = \frac{P(H_0)}{1 - P(H_0)}$ is the odds of H_0
- $H_1 = H_0 \cup \neg H_0$ recovers the classical formulation of the Bayesian theorem as $P(H_1 \mid D) = P(H_1) = 1$ and $P(D \mid H_1) = P(D)$

H: hypothesis
D: data
P: probability

prior "odds" of hypothesis H_0 versus H_1

posterior "odds" of hypothesis H_0 versus H_1
given the data D

Bayes factor

$$\underbrace{\frac{P(H_0 \mid D)}{P(H_1 \mid D)}}_{\text{posterior "odds"}} = \frac{P(H_0)}{P(H_1)} \times \frac{P(D \mid H_0)}{P(D \mid H_1)}$$

Special cases:

- $H_1 = \neg H_0$, then $\frac{P(H_0)}{P(H_1)} = \frac{P(H_0)}{P(\neg H_0)} = \frac{P(H_0)}{1 - P(H_0)}$ is the odds of H_0

H: hypothesis
D: data
P: probability

- $H_1 = H_0 \cup \neg H_0$ recovers the classical formulation of the Bayesian theorem, then $P(H_1) = 1 = P(H_1 \mid D)$ and $P(D \mid H_1) = P(D)$

$$\underbrace{\frac{P(H_0 \mid D)}{1}}_{\text{posterior probability}} = \frac{P(H_0)}{1} \times \frac{P(D \mid H_0)}{P(D)}$$

Special cases:

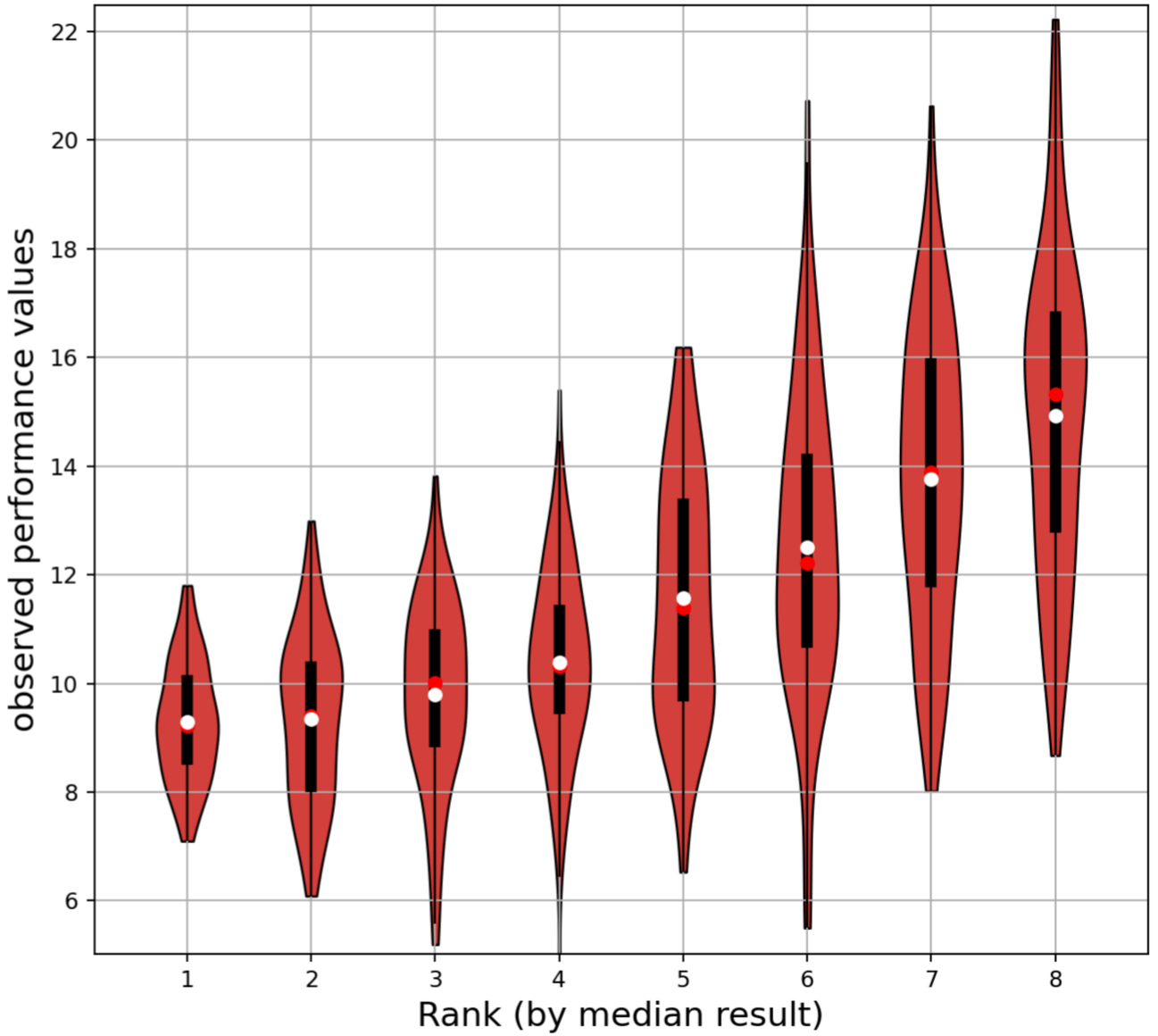
- $H_1 = \neg H_0$, then $\frac{P(H_0)}{P(H_1)} = \frac{P(H_0)}{P(\neg H_0)} = \frac{P(H_0)}{1 - P(H_0)}$ is the odds of H_0
- $H_1 = H_0 \cup \neg H_0$ recovers the classical formulation of the Bayesian theorem as $P(H_1 \mid D) = P(H_1) = 1$ and $P(D \mid H_1) = P(D)$

H: hypothesis
D: data
P: probability

Quantification of performance

quantify quantify quantify

What's wrong with ranking algorithms?



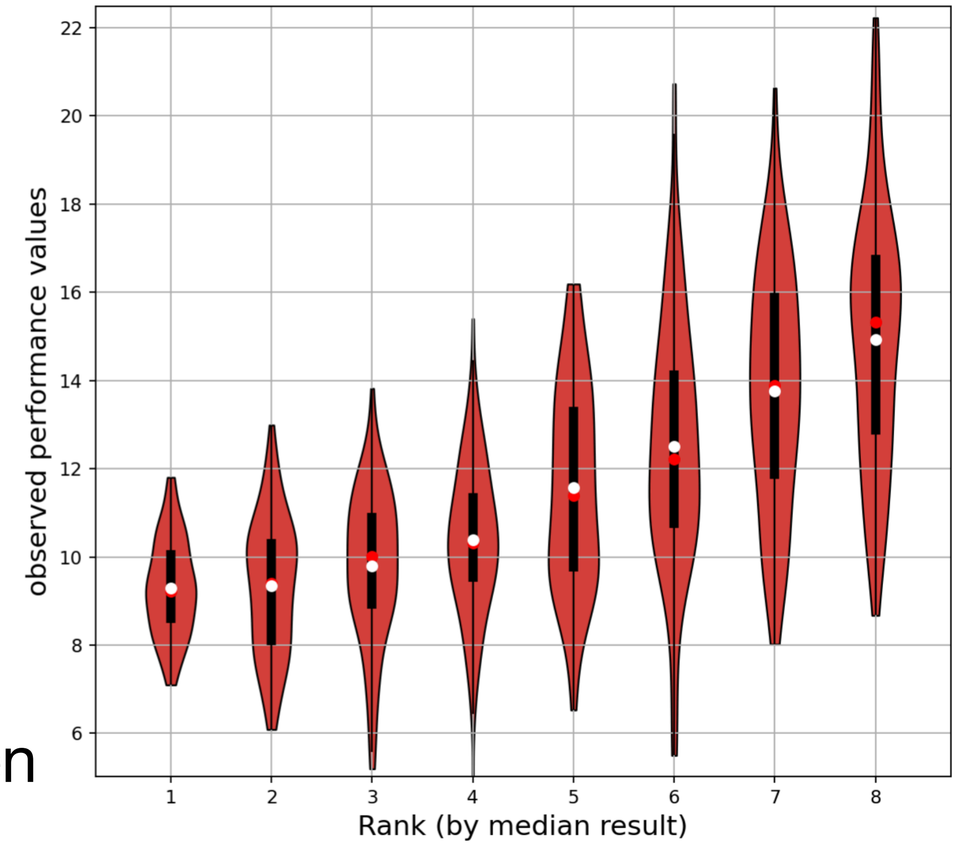
Inspired by Hoos 2023. Inconvenient Truths on Algorithm Competitions and Ways of Improving on Known Weaknesses. Presented at the Dagstuhl seminar 23251 *Challenges in Benchmarking Optimization Heuristics*.

What's wrong with ranking algorithms?

The ranking

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

- can not make a consistent distinction between (genuinely) equal and non-equal ranks



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

Four Levels of Measurement (Scales)

- Nominal scale - categorial, define a classification
- Ordinal scale - define an order, e.g., **ranks**, function values (arguably)
- Interval scale - differences are quantitatively meaningful
- Ratio scale - 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).

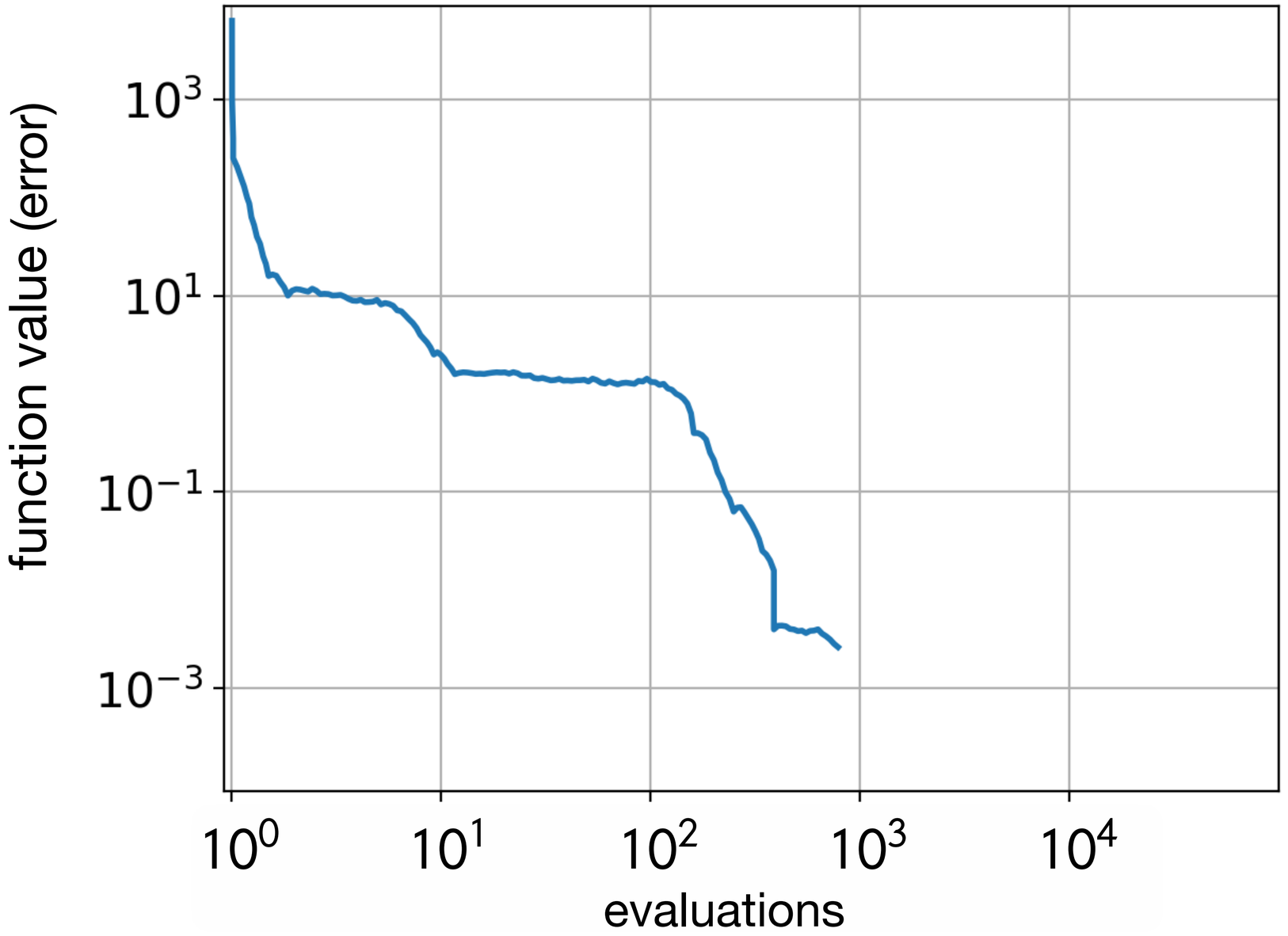
Empirical Cumulative Distributions

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.

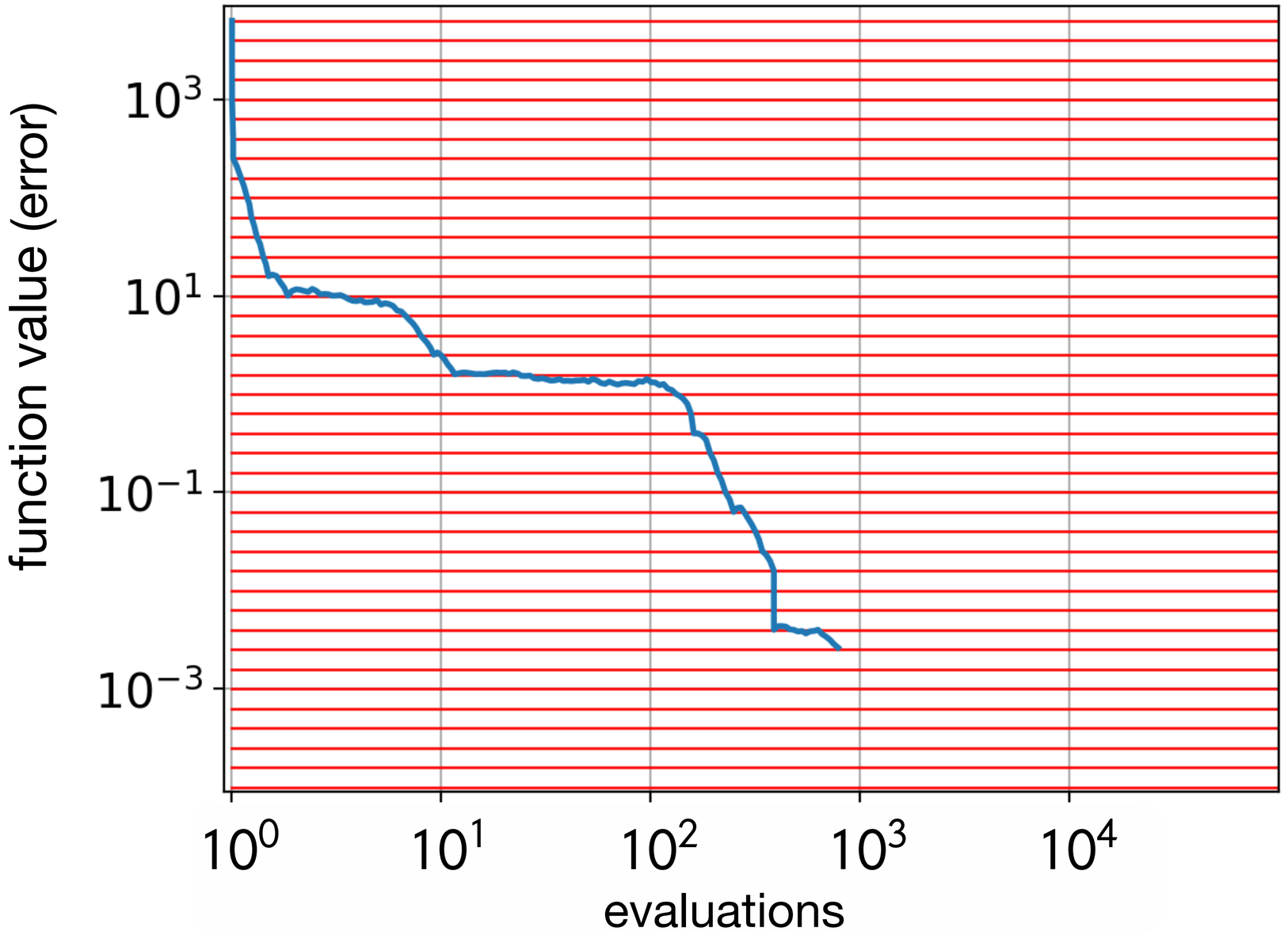
Hansen et al. 2021. A platform for comparing continuous optimizers in a black-box setting. *Optimization Methods and Software*, *36*(1).

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).

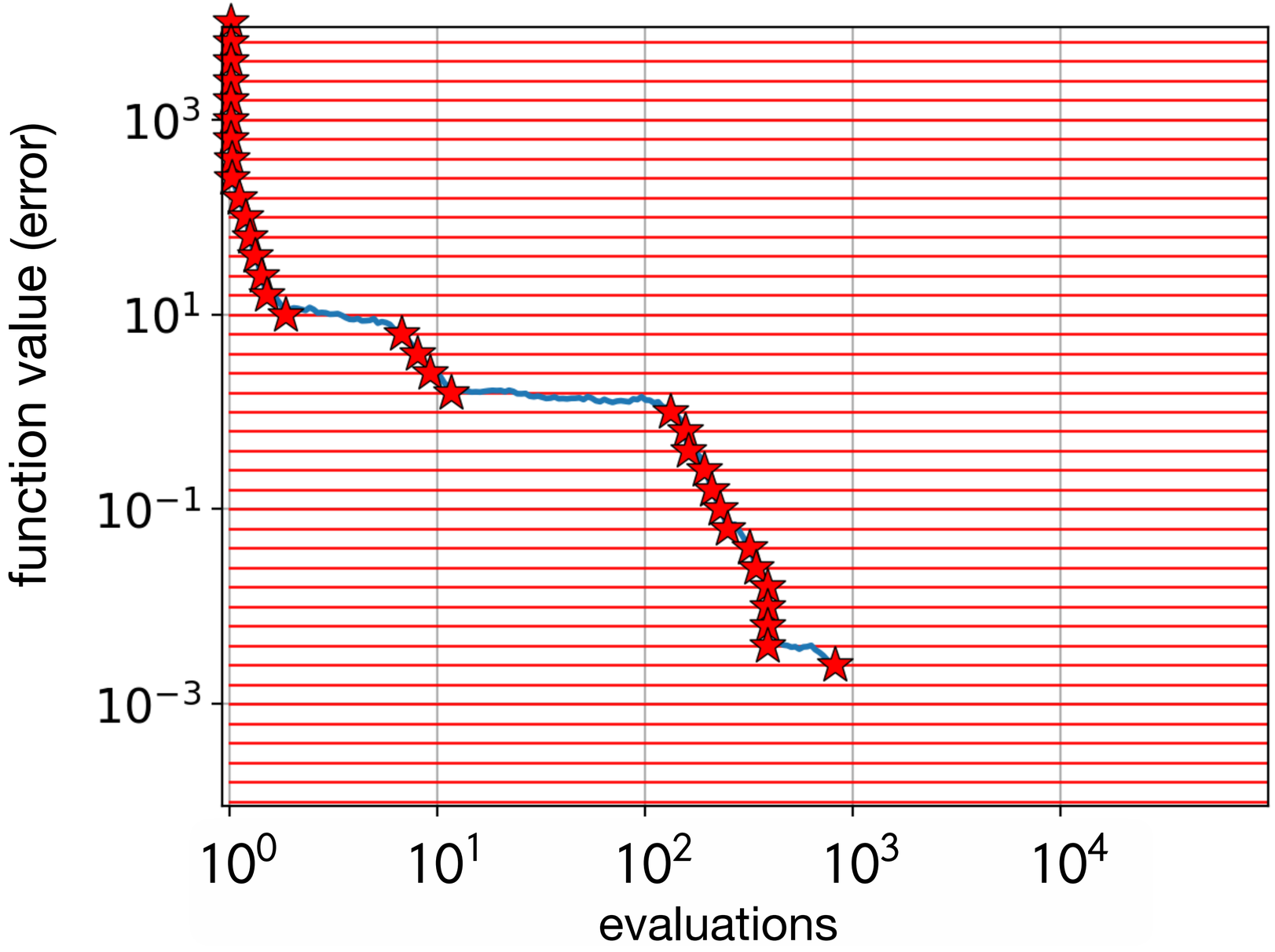
From a Convergence Graph to the Empirical Runtime Distribution



equidistance
“target” values

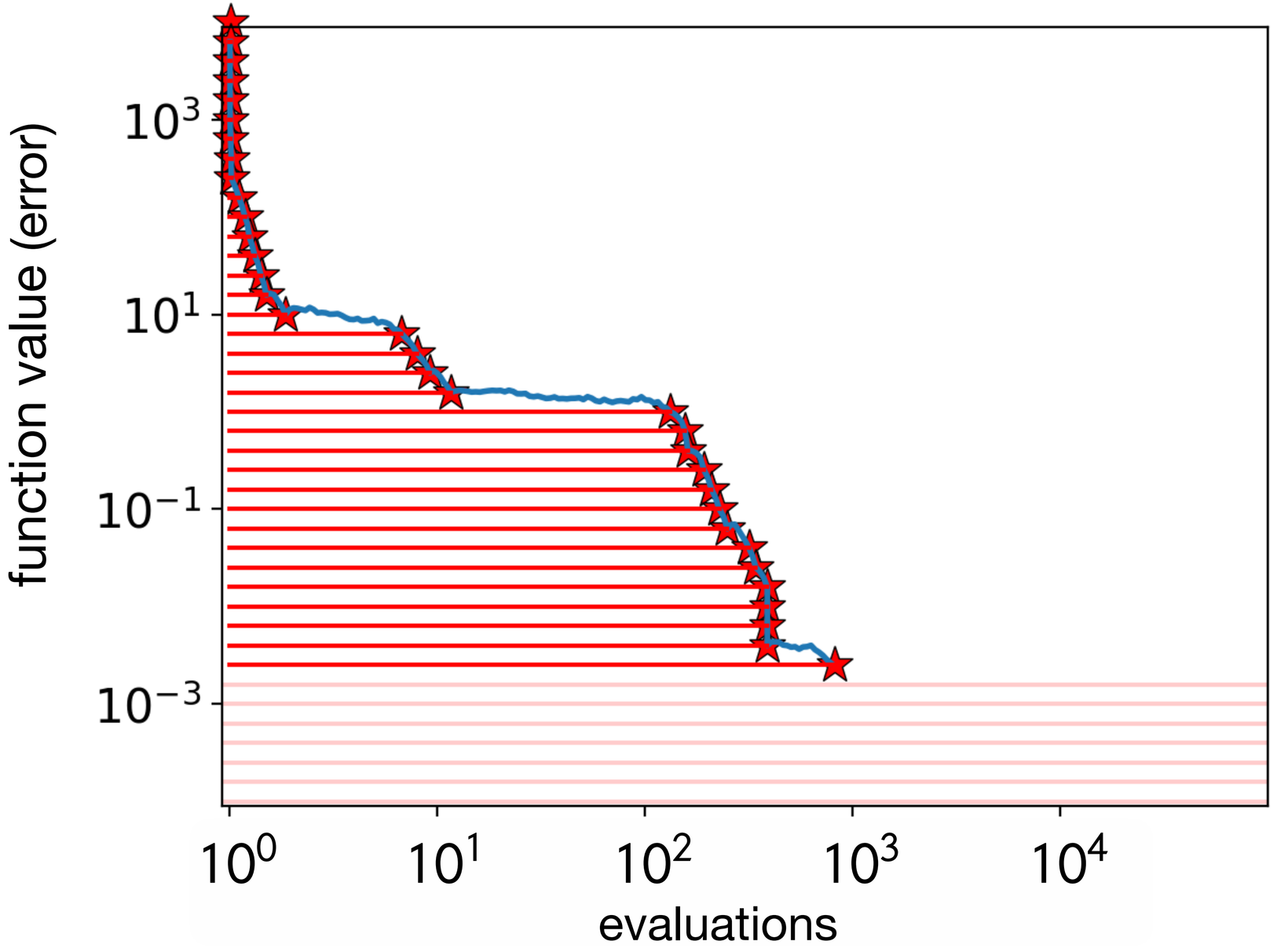
Hansen et al. 2022. Anytime performance assessment in blackbox optimization benchmarking. *IEEE Trans. on EC*, 26(6).

From a Convergence Graph to the Empirical Runtime Distribution



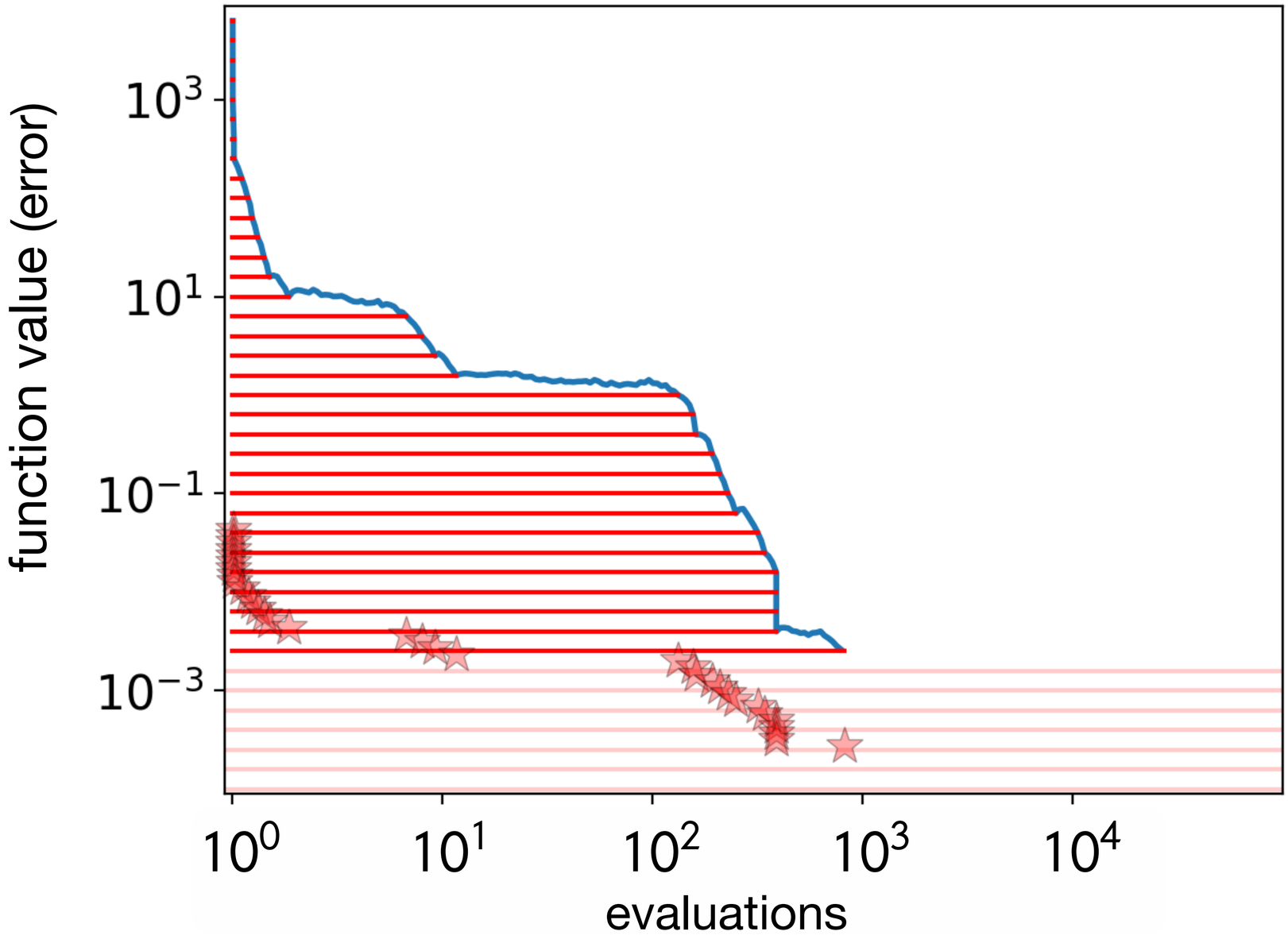
Hansen et al. 2022. Anytime performance assessment in blackbox optimization benchmarking. *IEEE Trans. on EC*, 26(6).

From a Convergence Graph to the Empirical Runtime Distribution



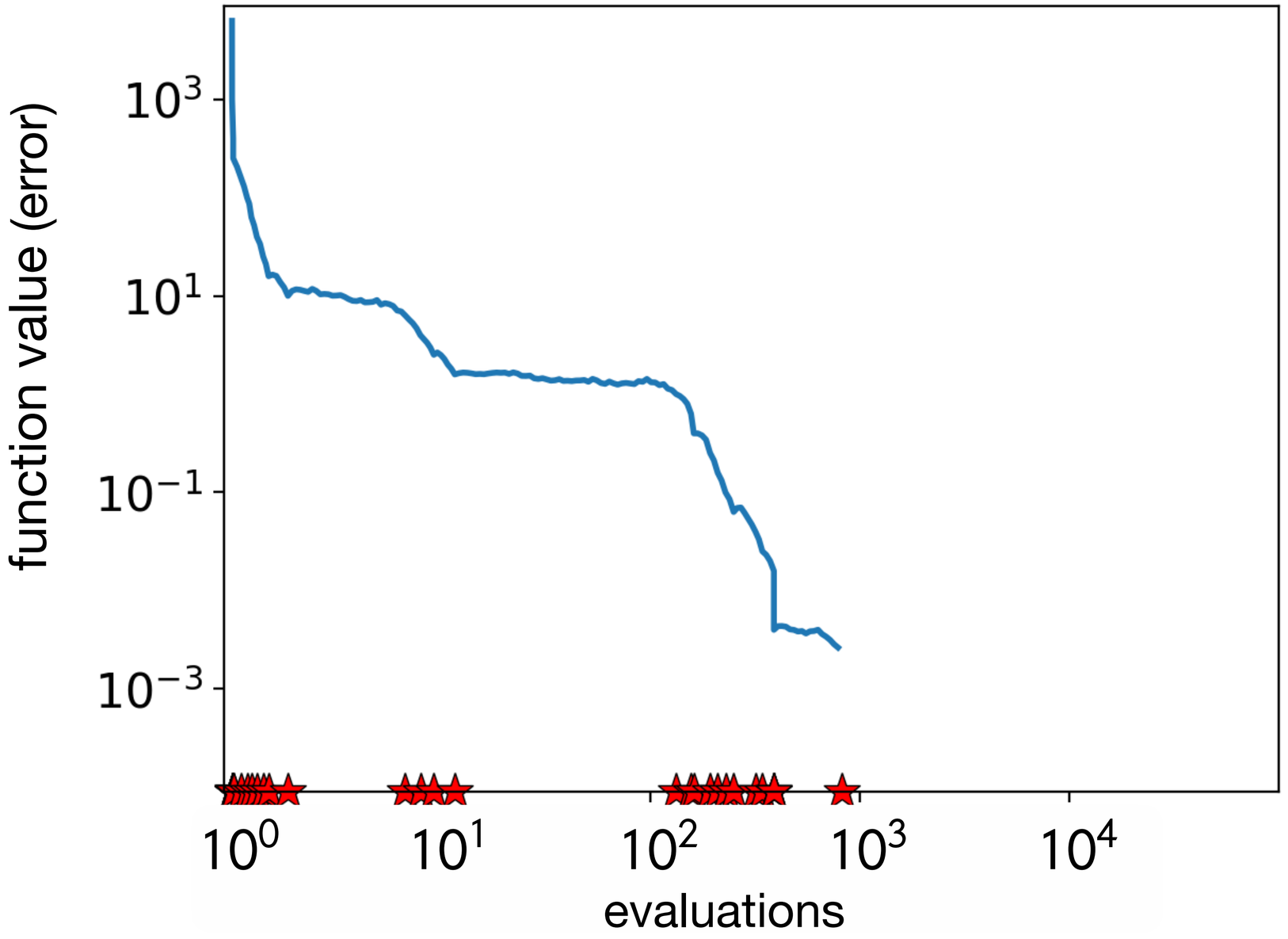
Hansen et al. 2022. Anytime performance assessment in blackbox optimization benchmarking. *IEEE Trans. on EC*, 26(6).

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).

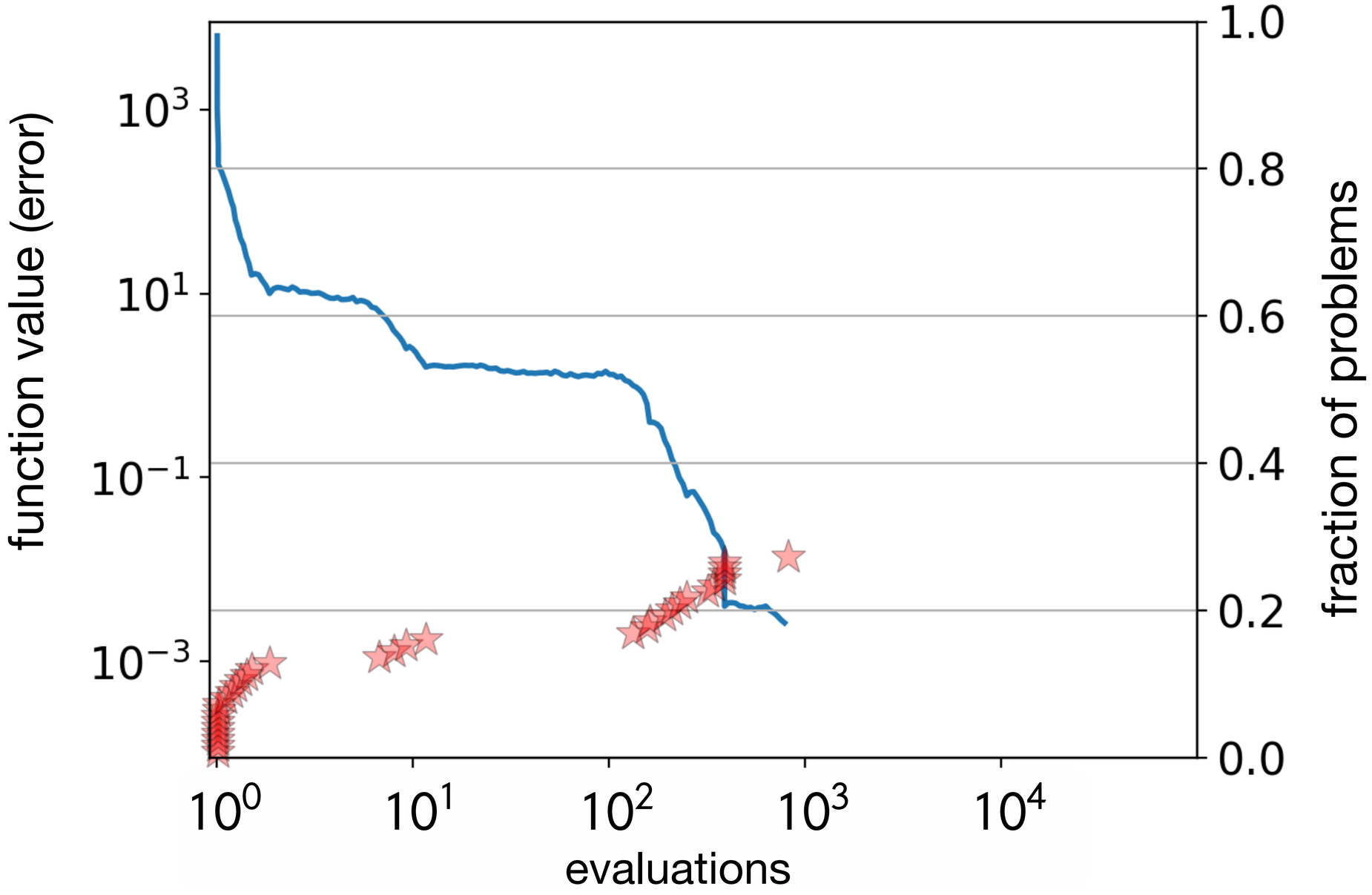
From a Convergence Graph to the Empirical Runtime Distribution



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

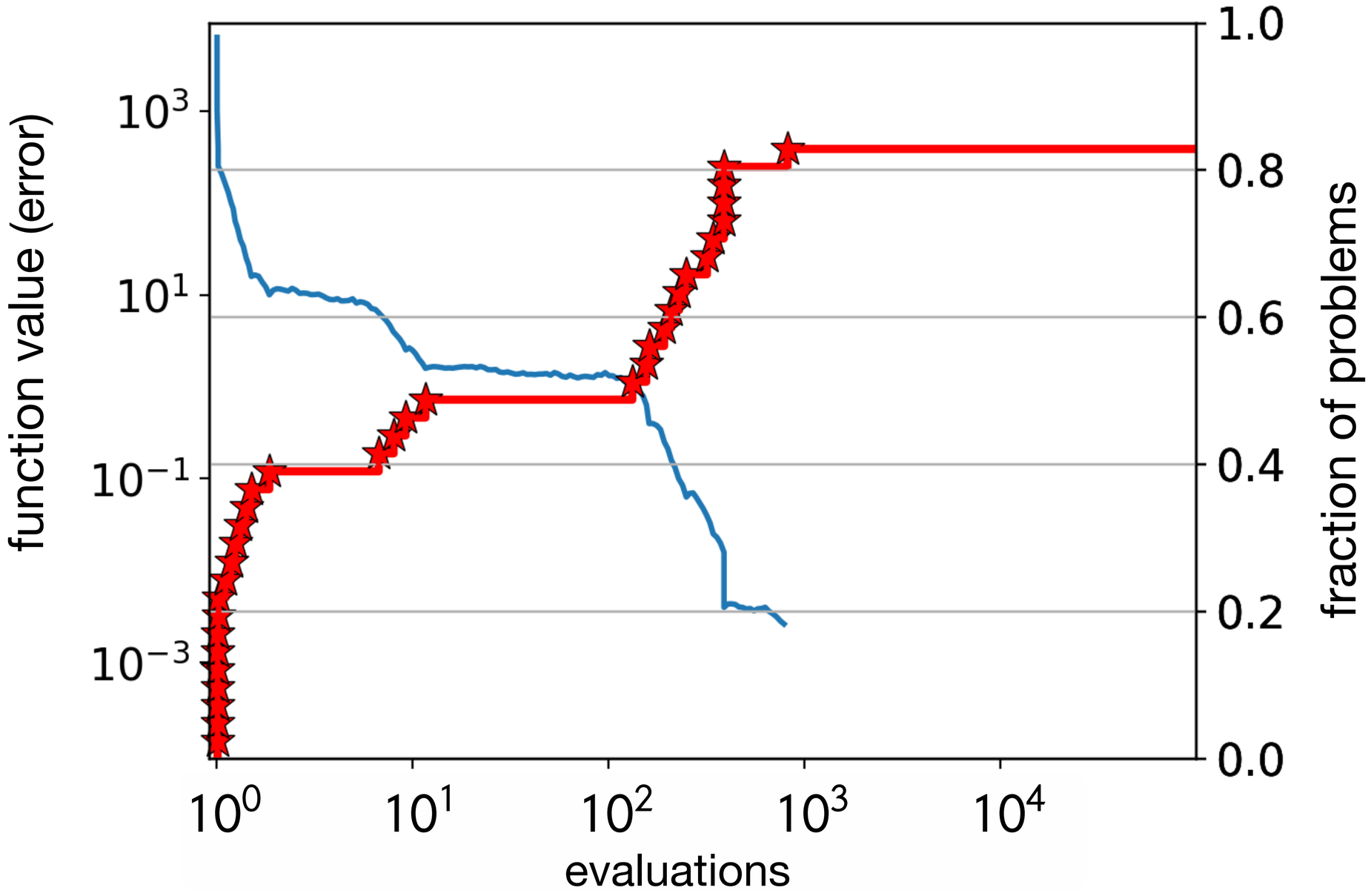
Hansen et al. 2022. Anytime performance assessment in blackbox optimization benchmarking. *IEEE Trans. on EC*, 26(6).

From a Convergence Graph to the Empirical Runtime Distribution



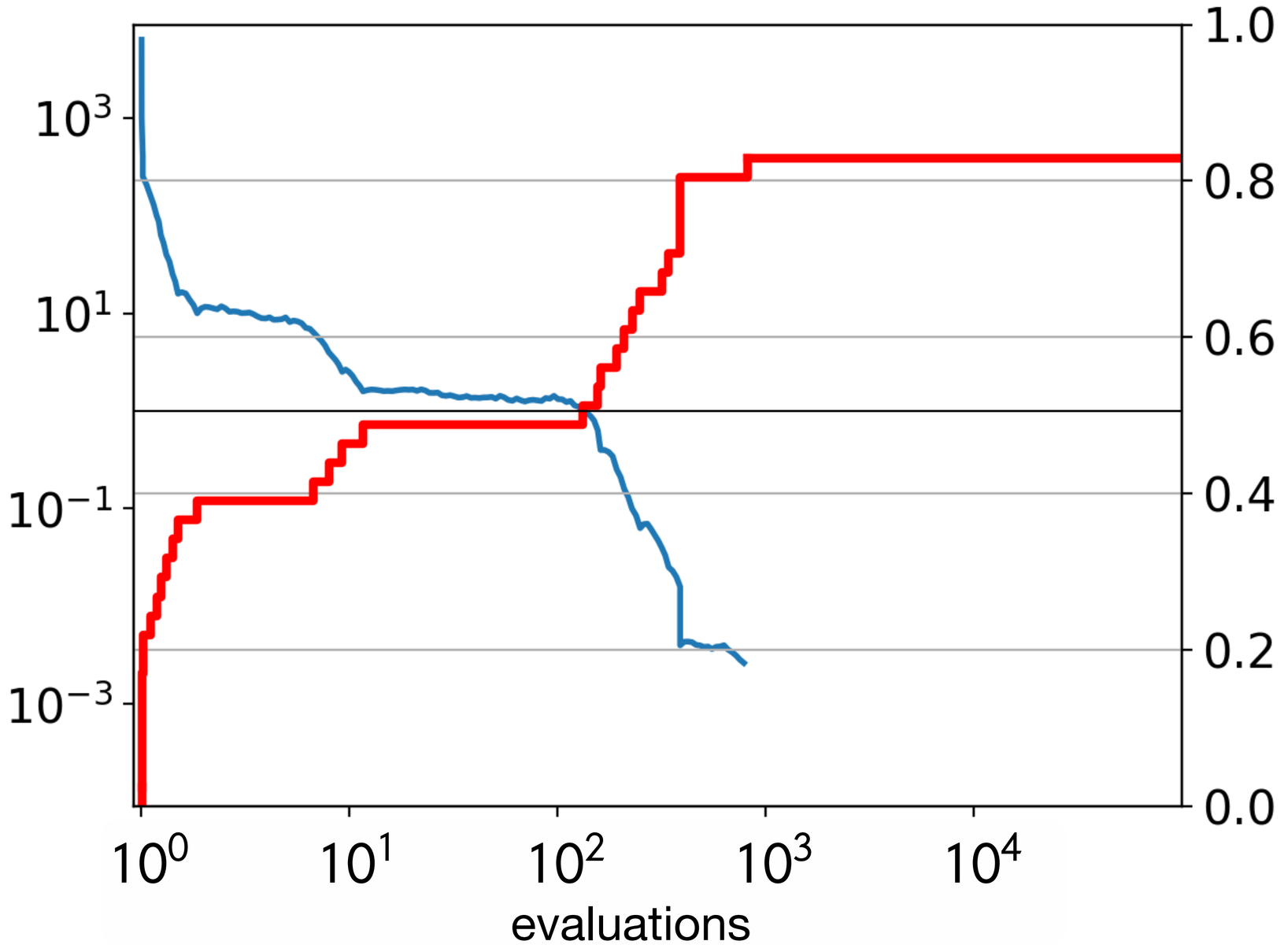
Hansen et al. 2022. Anytime performance assessment in blackbox optimization benchmarking. *IEEE Trans. on EC*, 26(6).

From a Convergence Graph to the Empirical Runtime Distribution



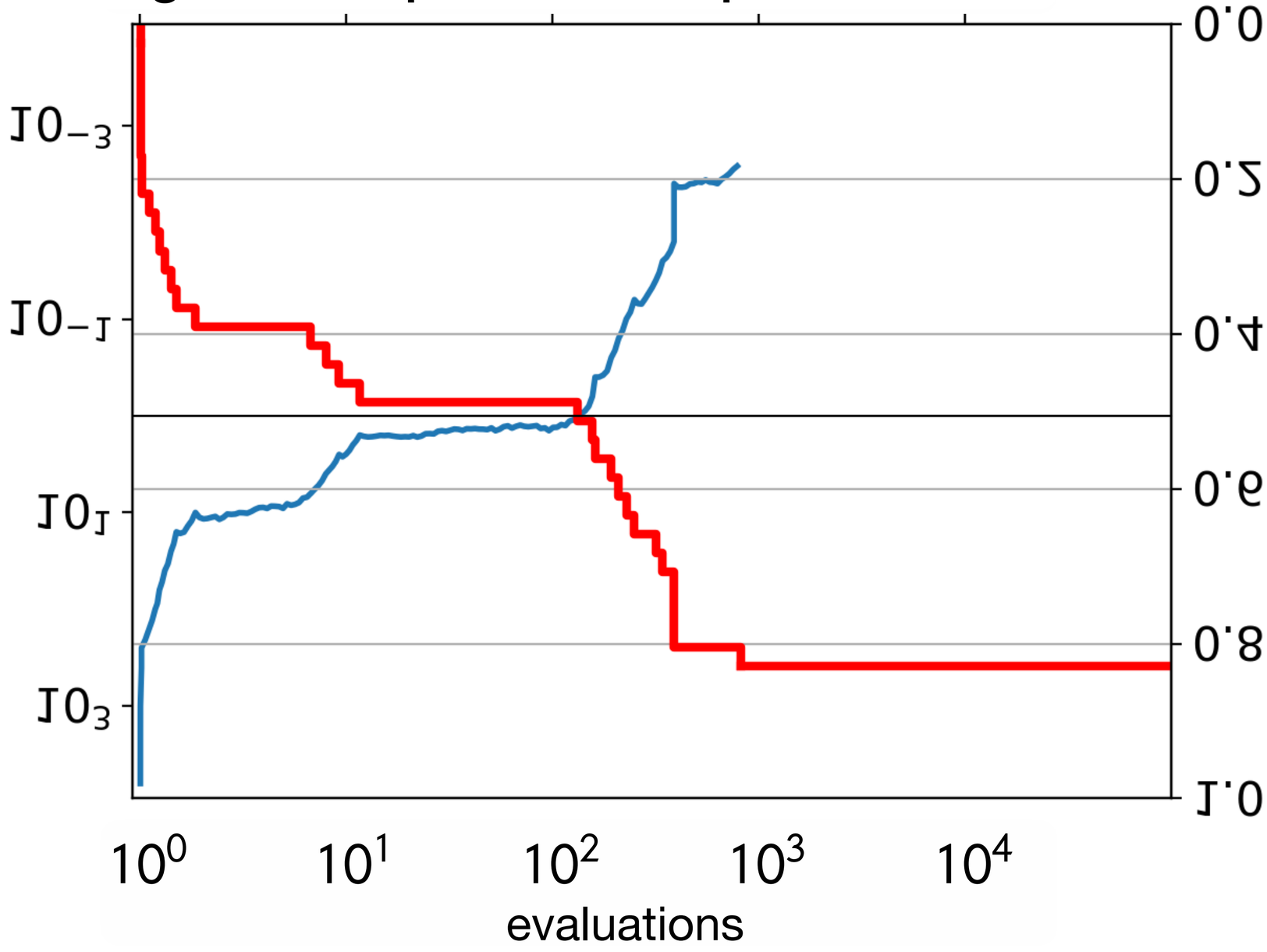
Hansen et al. 2022. Anytime performance assessment in blackbox optimization benchmarking. *IEEE Trans. on EC*, 26(6).

From a Convergence Graph to the Empirical Runtime Distribution



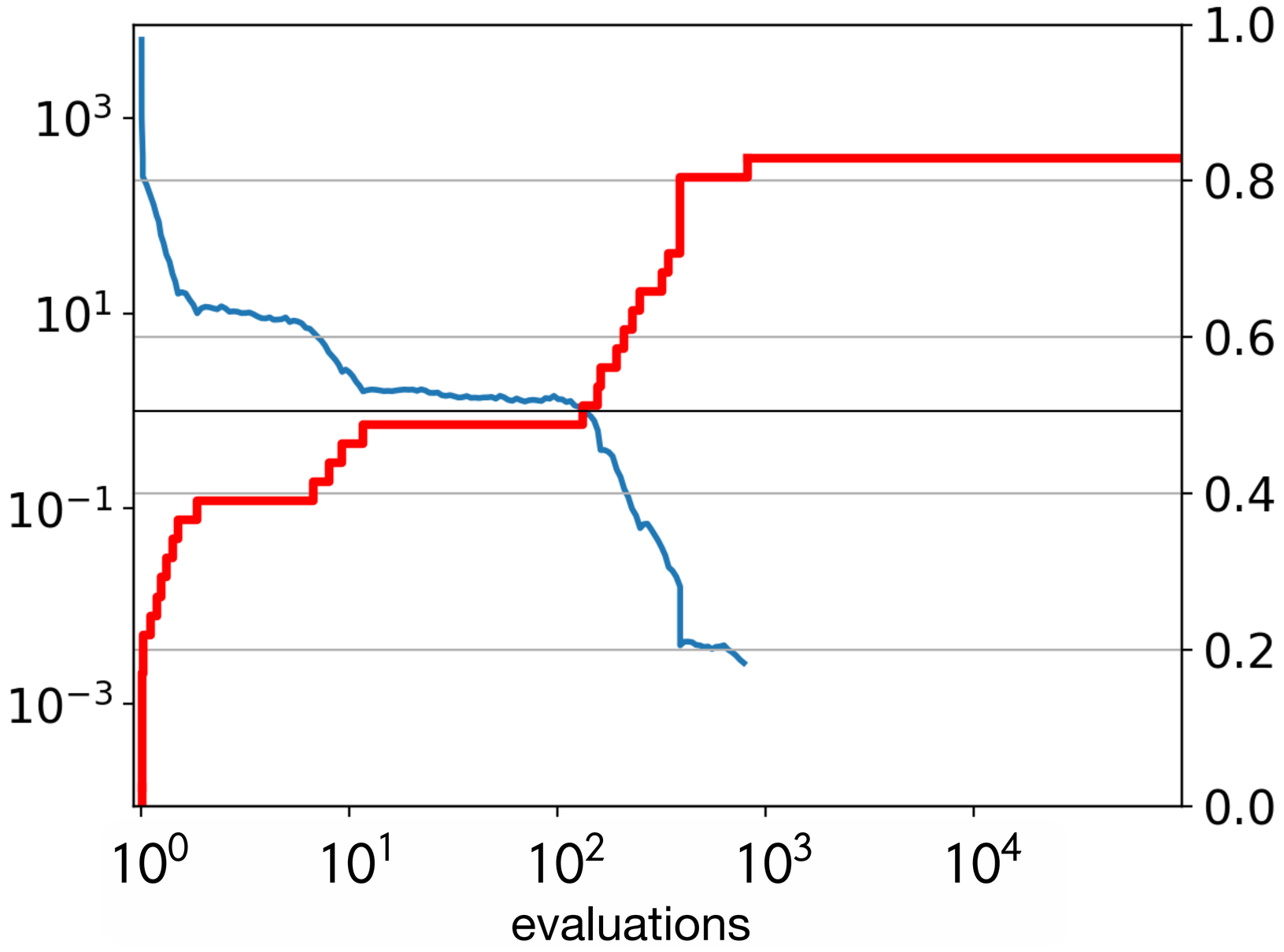
Hansen et al. 2022. Anytime performance assessment in blackbox optimization benchmarking. *IEEE Trans. on EC*, 26(6).

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).

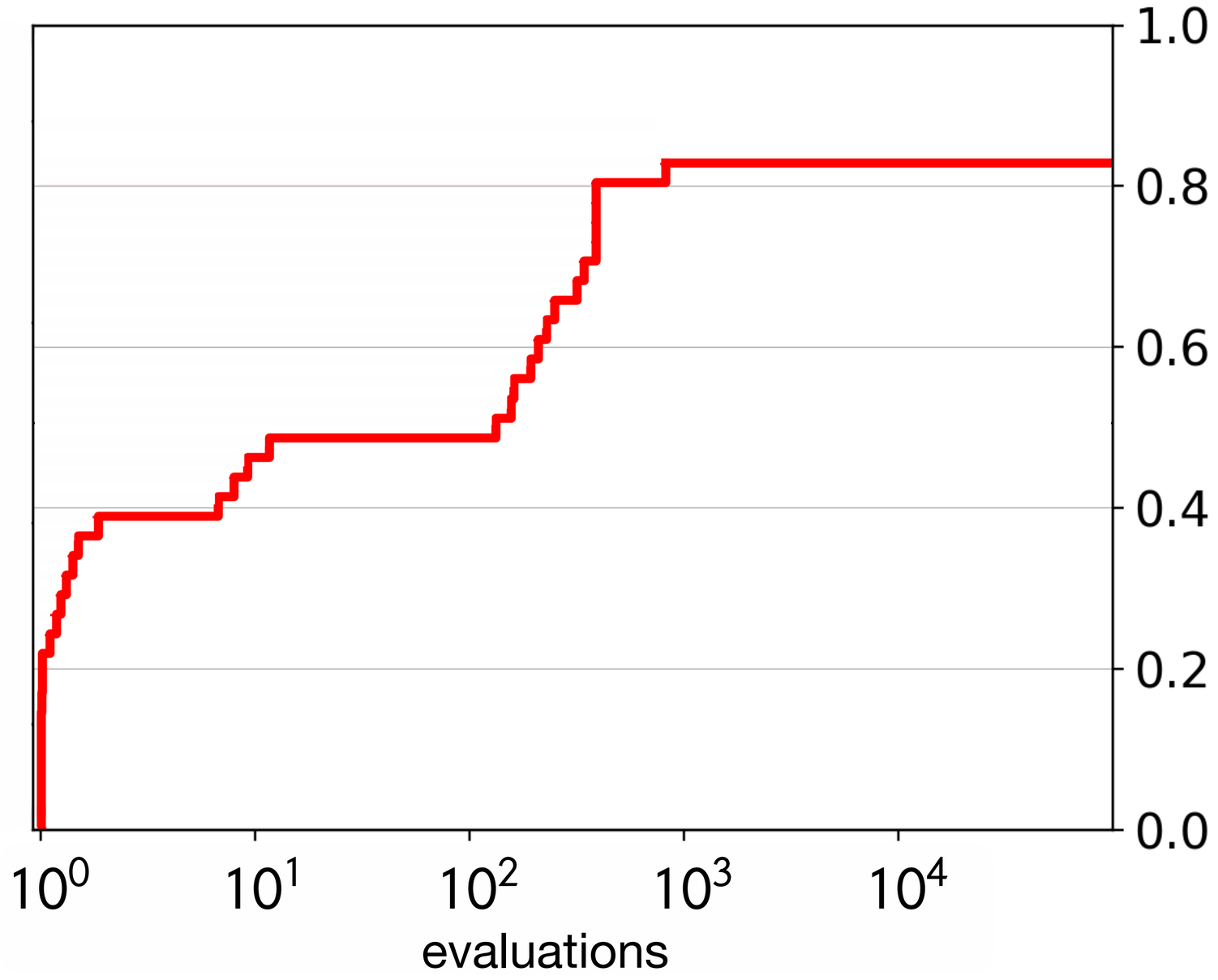
From a Convergence Graph to the Empirical Runtime Distribution



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*



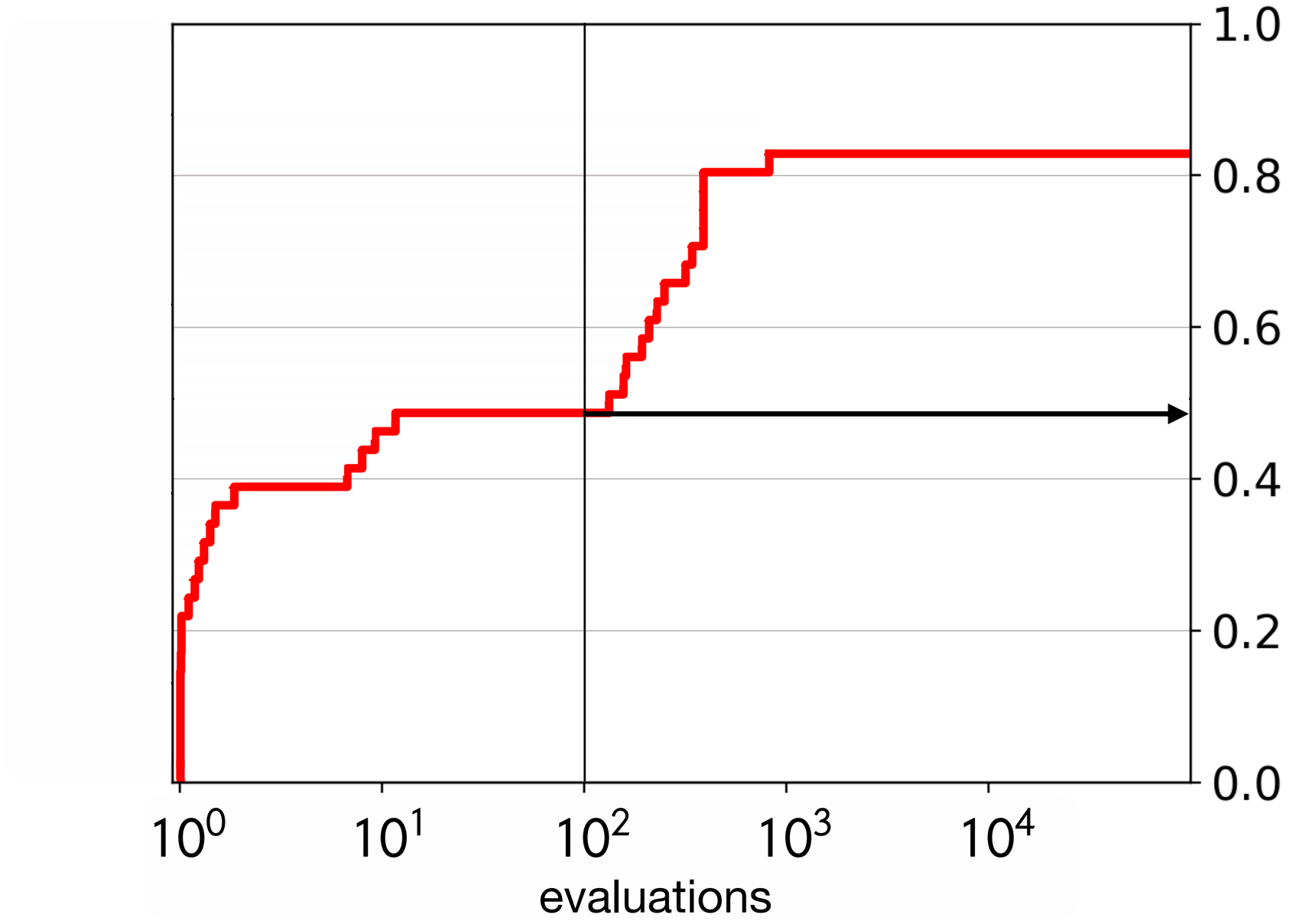
Hansen et al. 2022. Anytime performance assessment in blackbox optimization benchmarking. *IEEE Trans. on EC*, 26(6).

How to read Runtime Distribution graphs

- for any given budget (x-value): an observed success probability or attainment ratio
- for any given attainment ratio (y-value):
  - an observed required budget
  - an average runtime

From a Convergence Graph to the Empirical Runtime Distribution

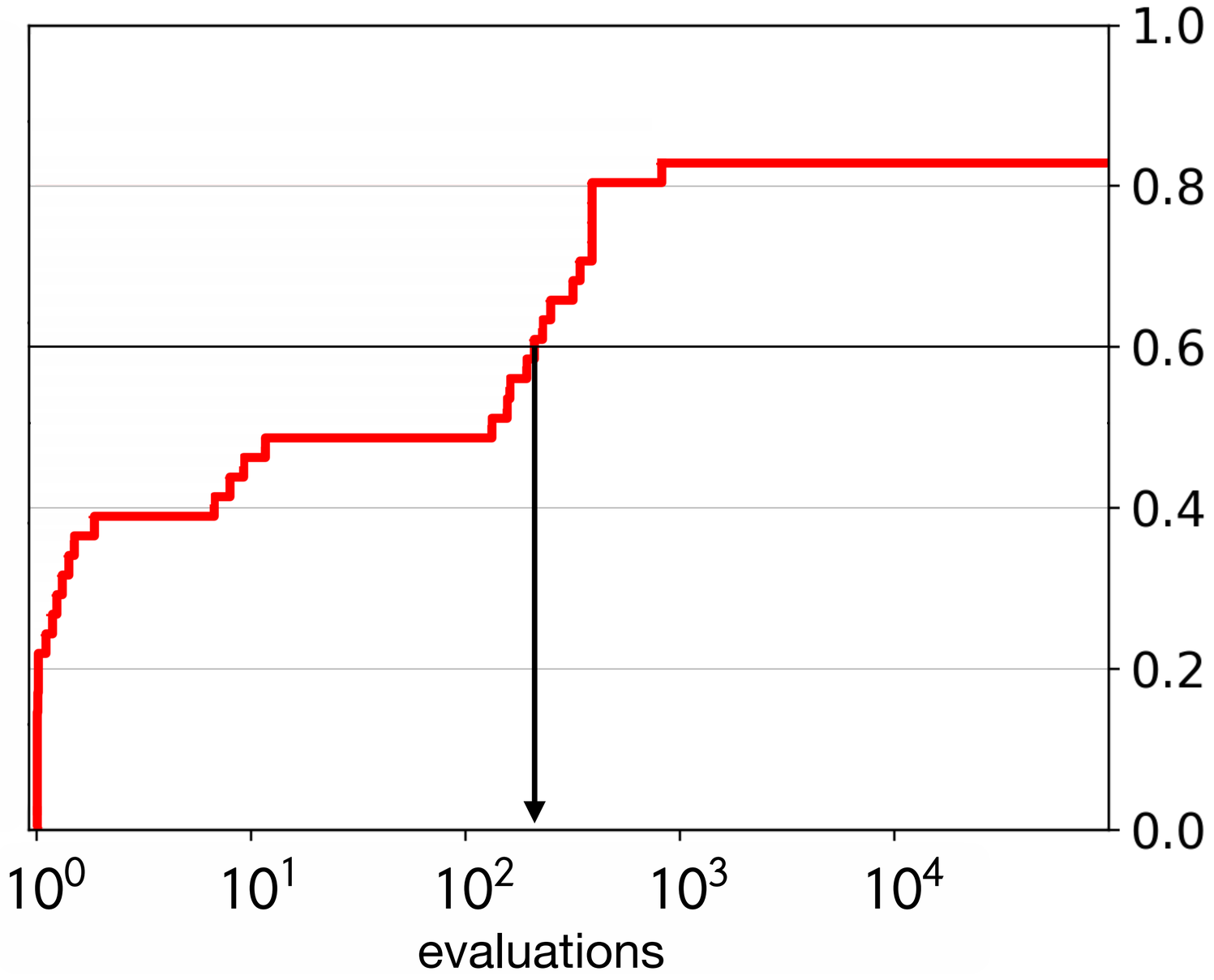
arrow tip: success probability or attainment for a budget of 100



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

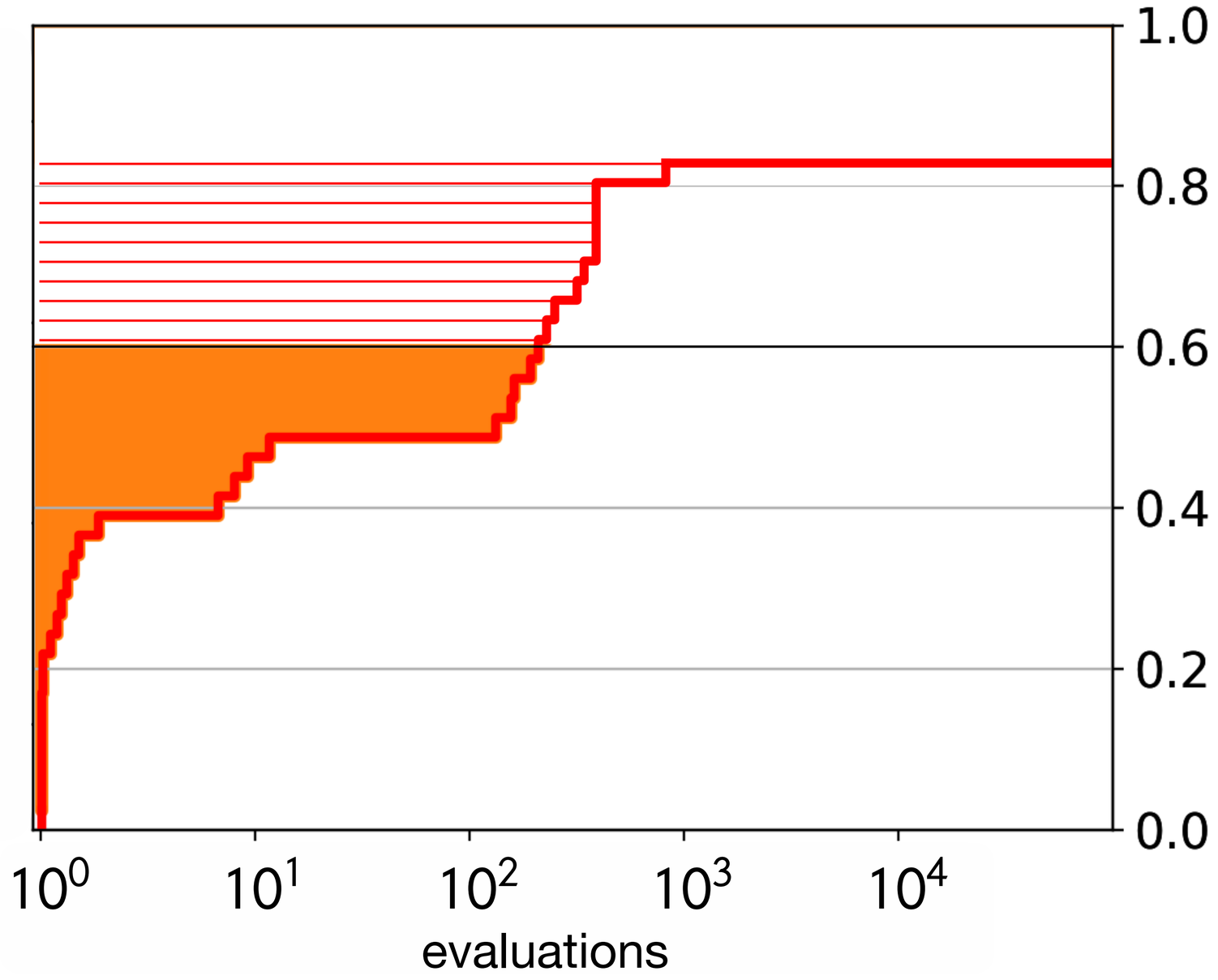
arrow tip: runtime of the 60% most difficult function or problem



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

orange area: average runtime of the 60% easiest functions or problems

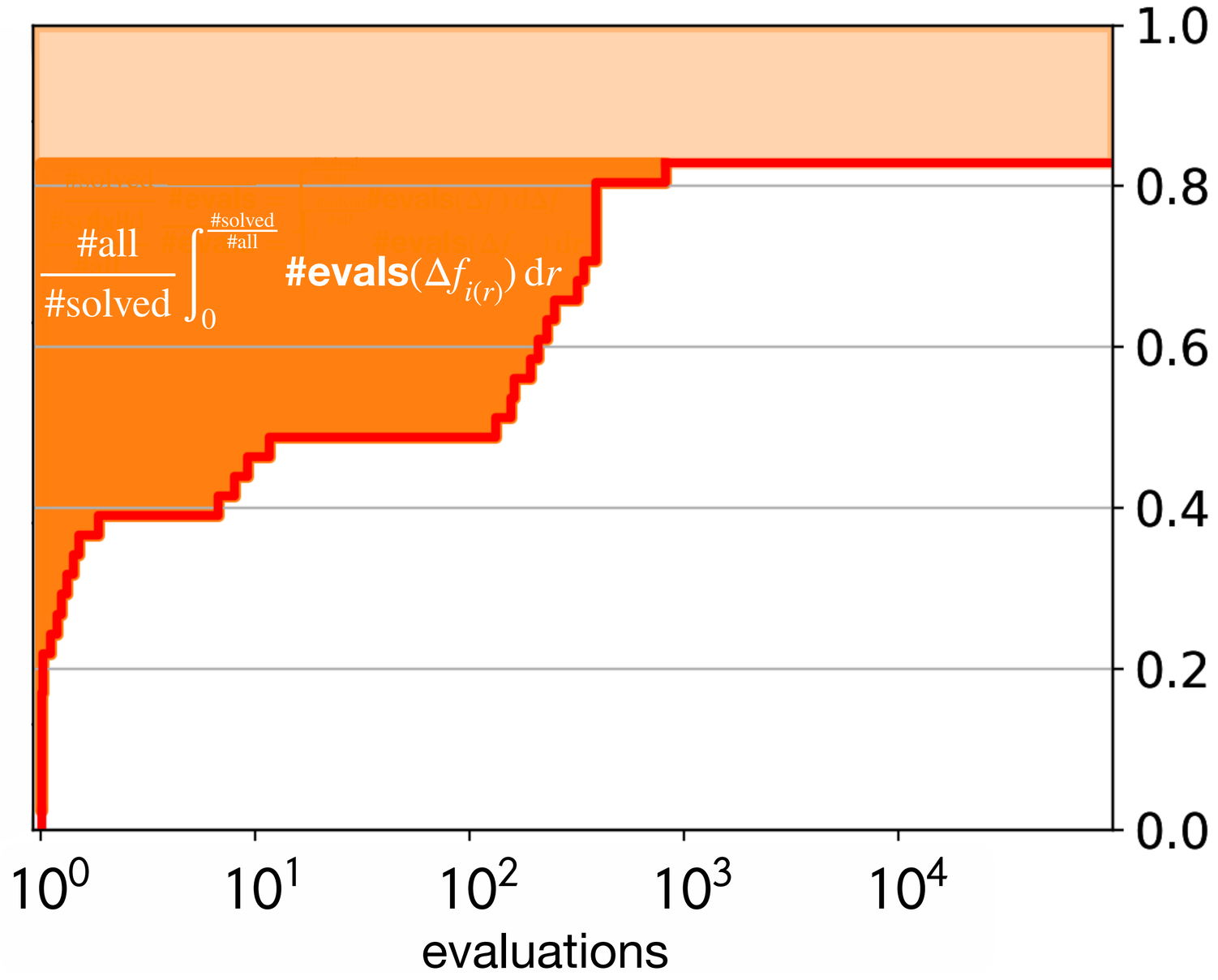


Hansen et al. 2022. Anytime performance assessment in blackbox optimization benchmarking. *IEEE Trans. on EC*, 26(6).

Empirical Runtime Distribution and area above the curve

the average **width** of the orange area represents the average runtime *of successful runs*

When the x-axis is in log-scale, the area is the *geometric* average

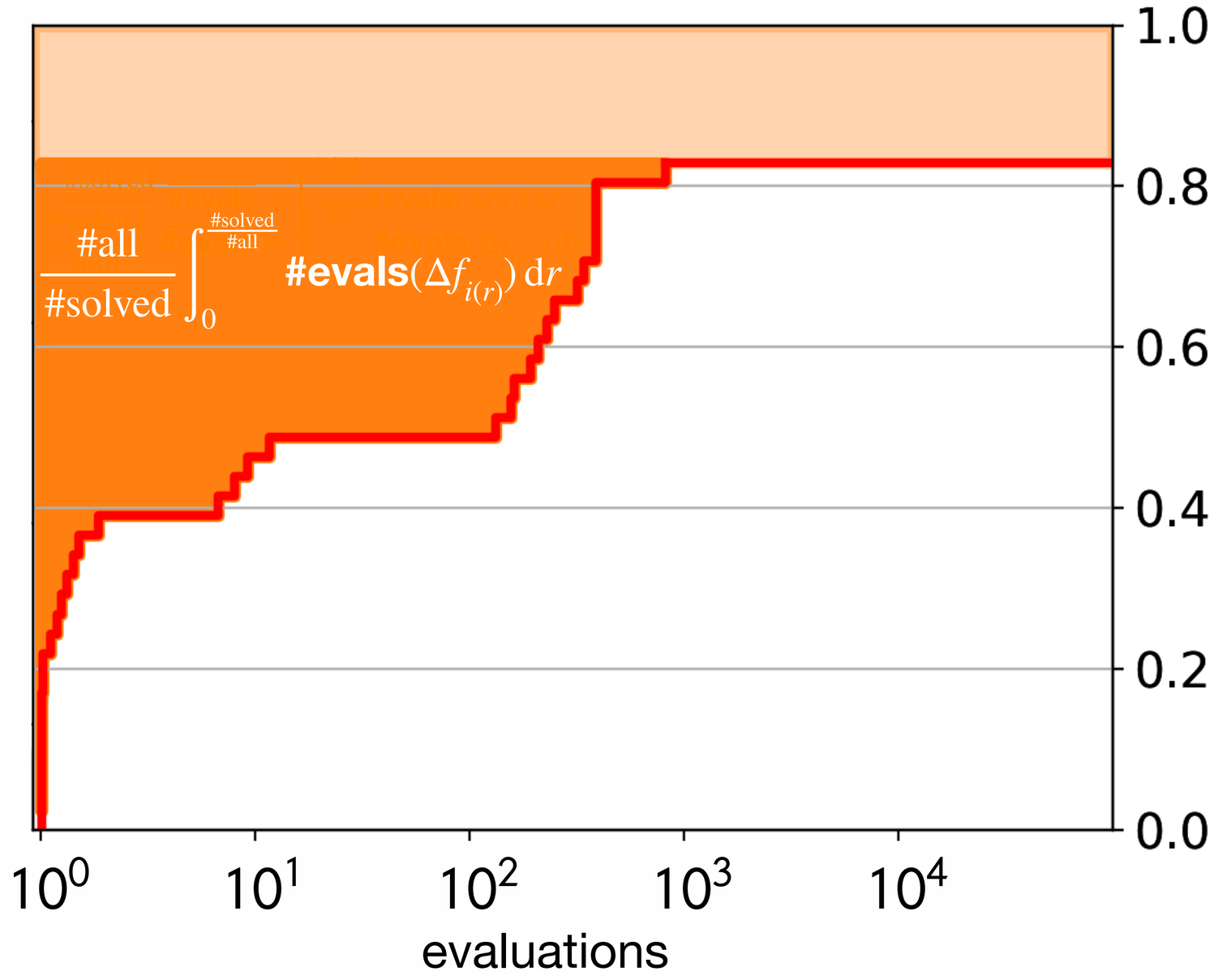


Hansen et al. 2022. Anytime performance assessment in blackbox optimization benchmarking. *IEEE Trans. on EC*, 26(6).

Empirical Runtime Distribution and area above the curve

the average **width** of the orange area represents the average runtime *of successful runs*

When the x-axis is in log-scale, the area is the *geometric* average



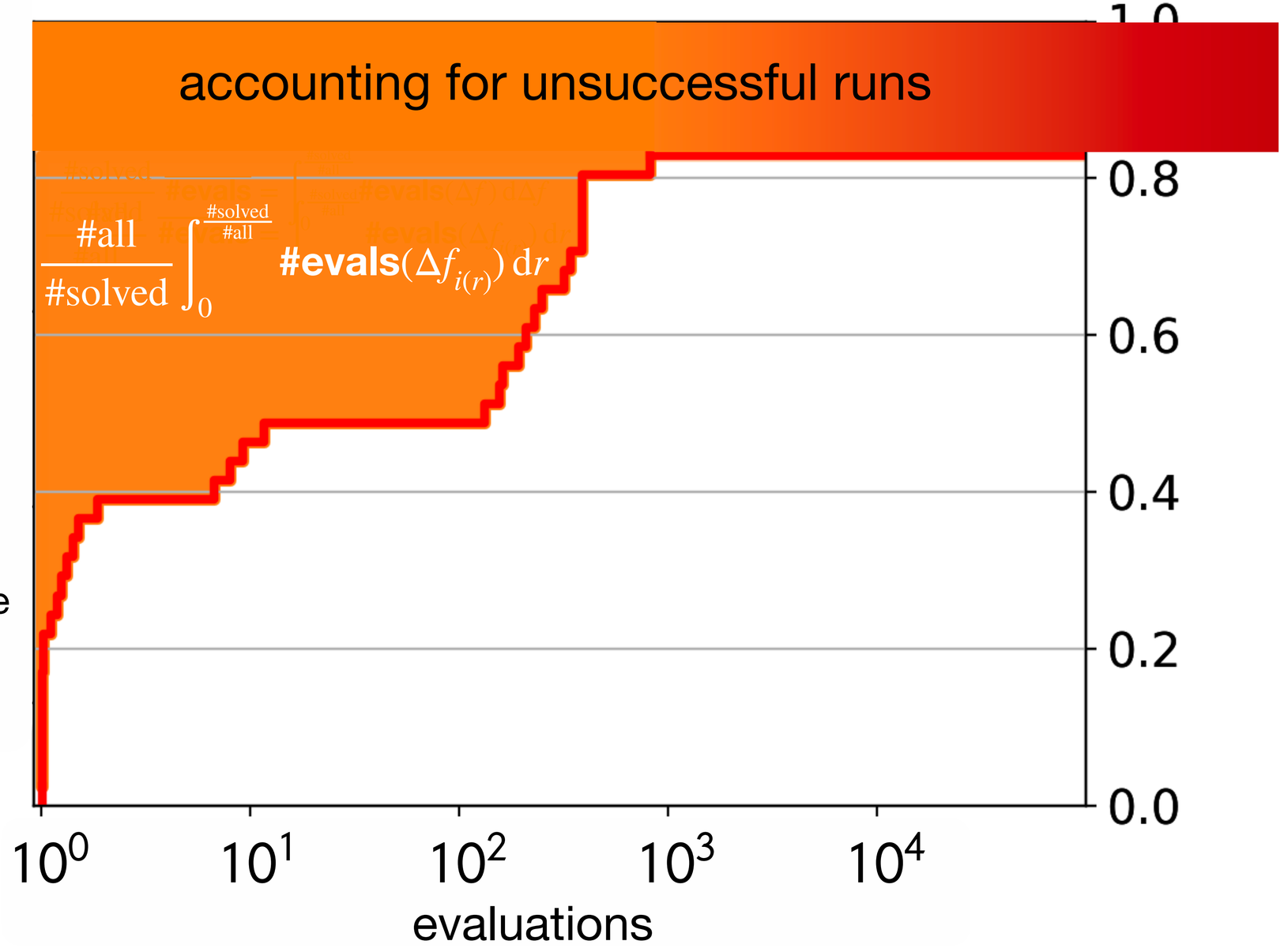
Hansen et al. 2022. Anytime performance assessment in blackbox optimization benchmarking. *IEEE Trans. on EC*, 26(6).

Empirical Runtime Distribution and area above the curve

the average **width** of the orange area represents the average runtime *of successful runs*

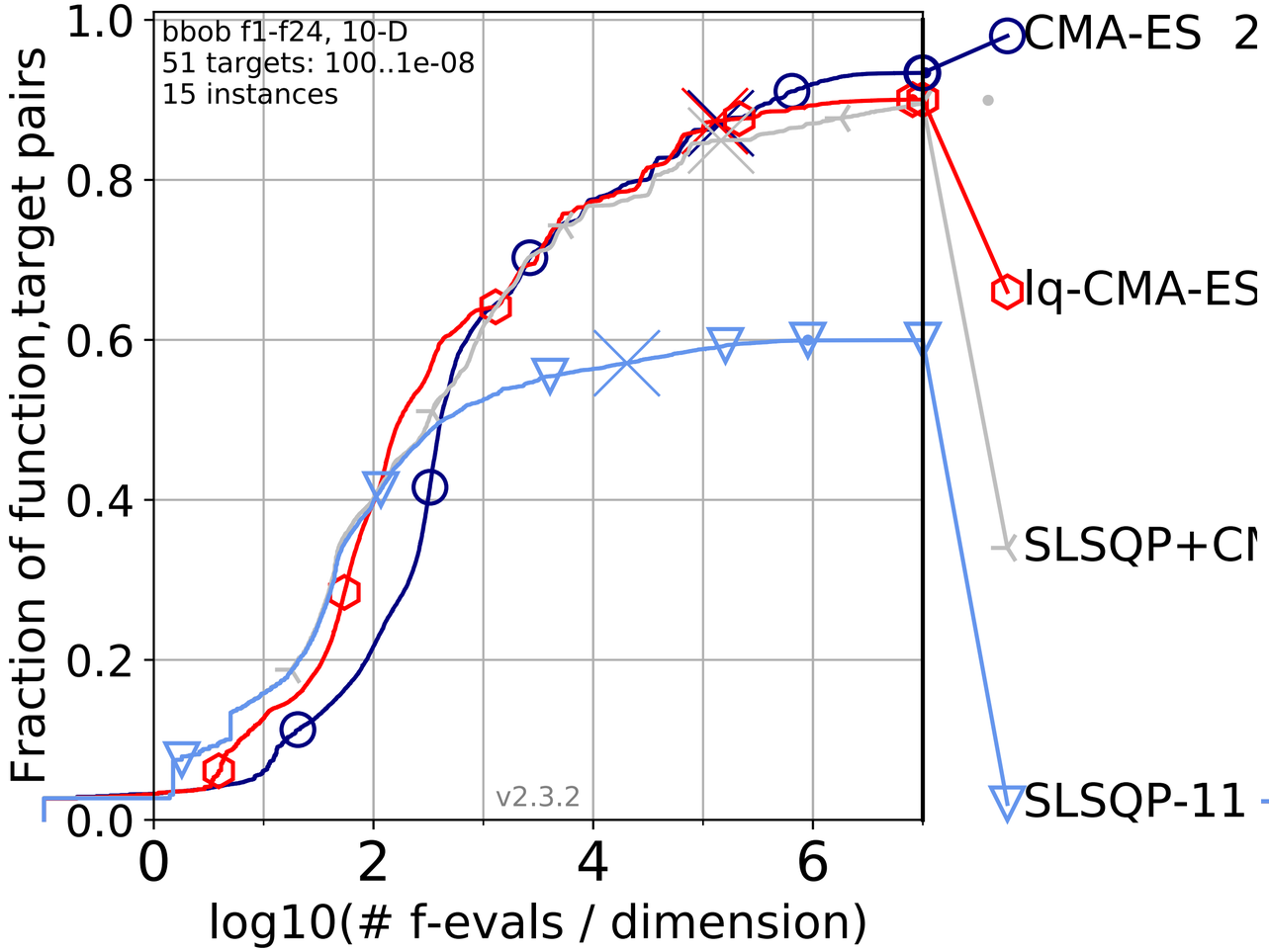
When the x-axis is in log-scale, the area is the *geometric* average

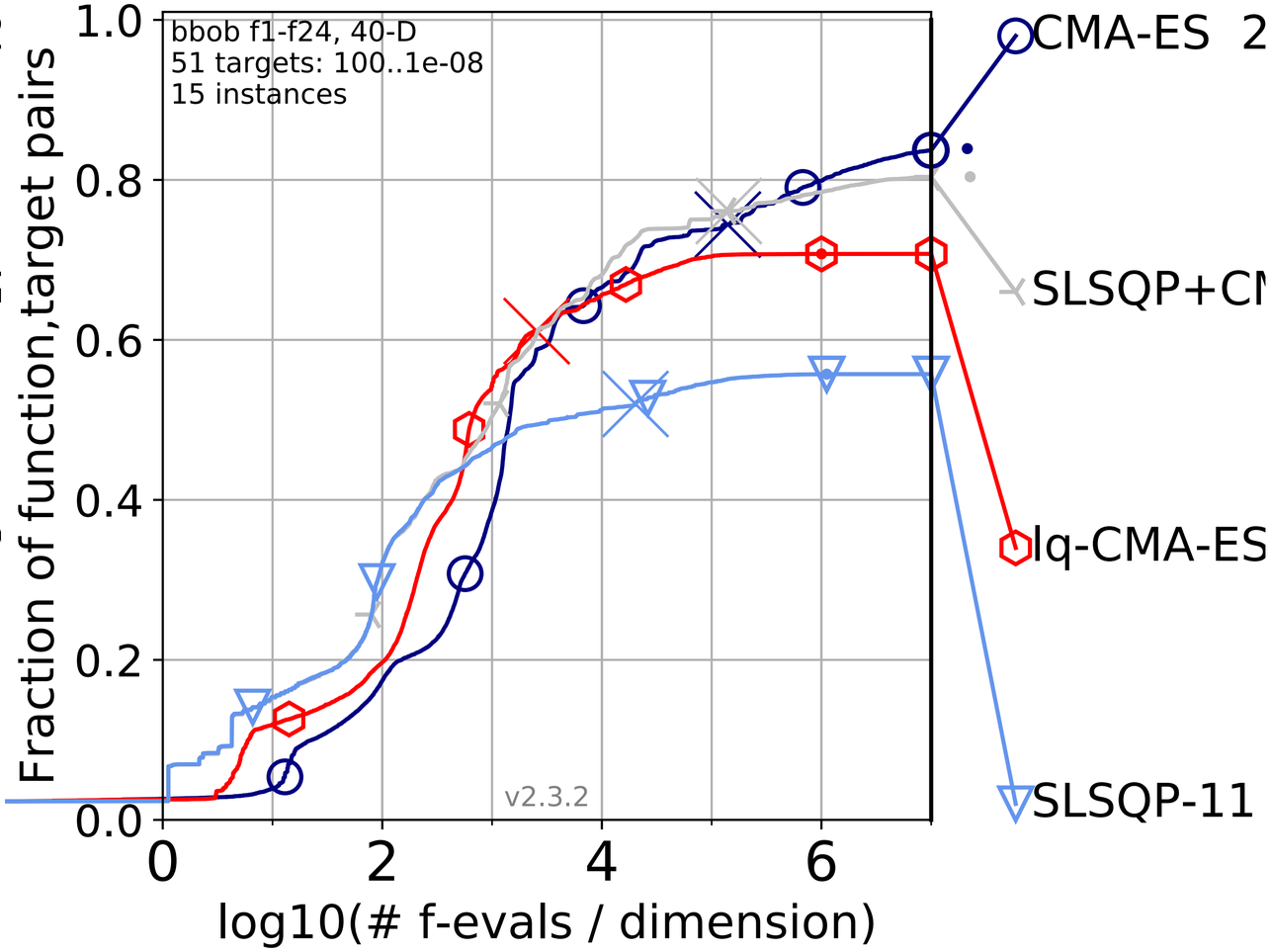
unsuccessful runs add runtime without solving the problem. Ideally, we measure their runtime too.



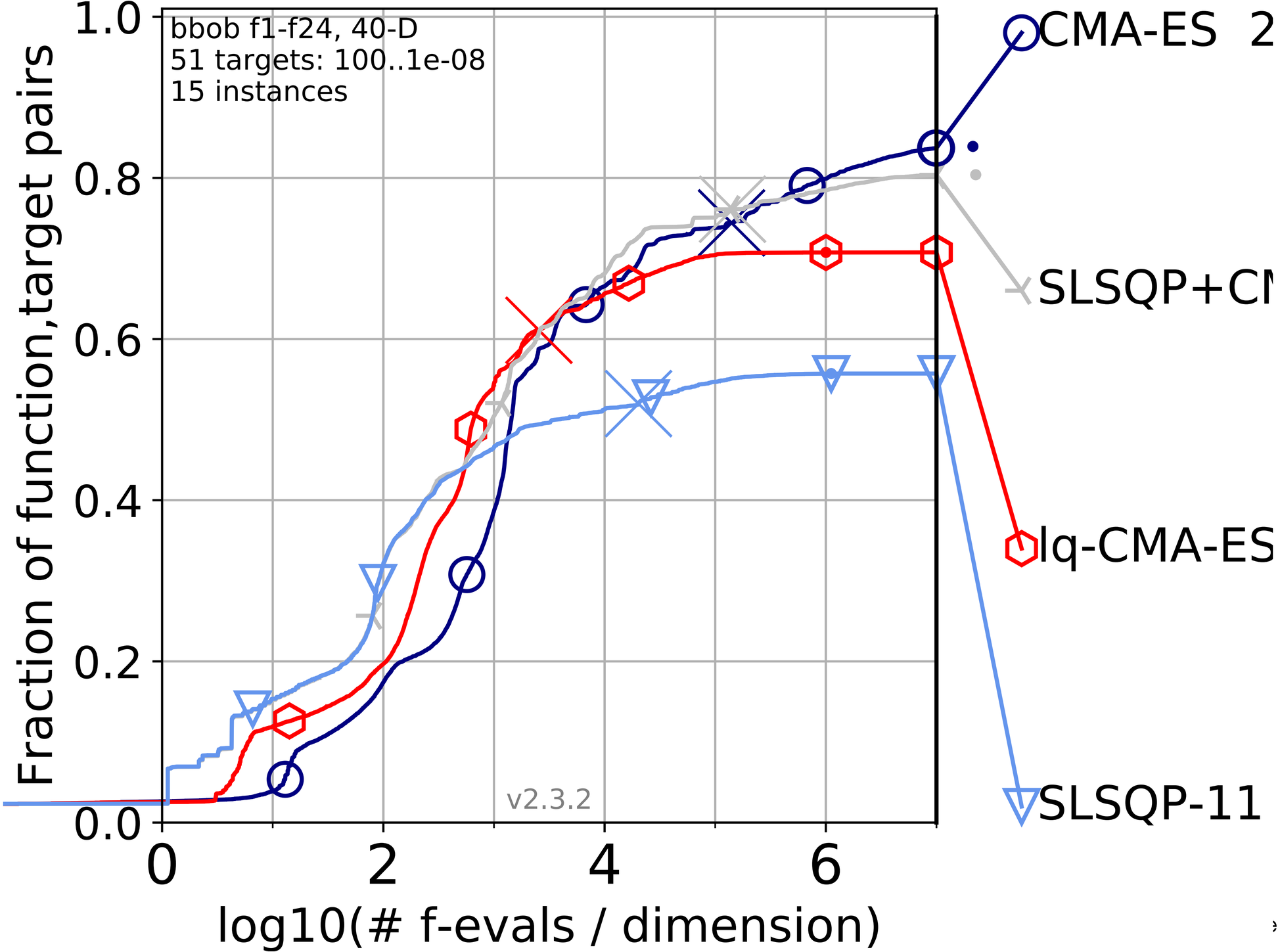
Hansen et al. 2022. Anytime performance assessment in blackbox optimization benchmarking. *IEEE Trans. on EC*, 26(6).

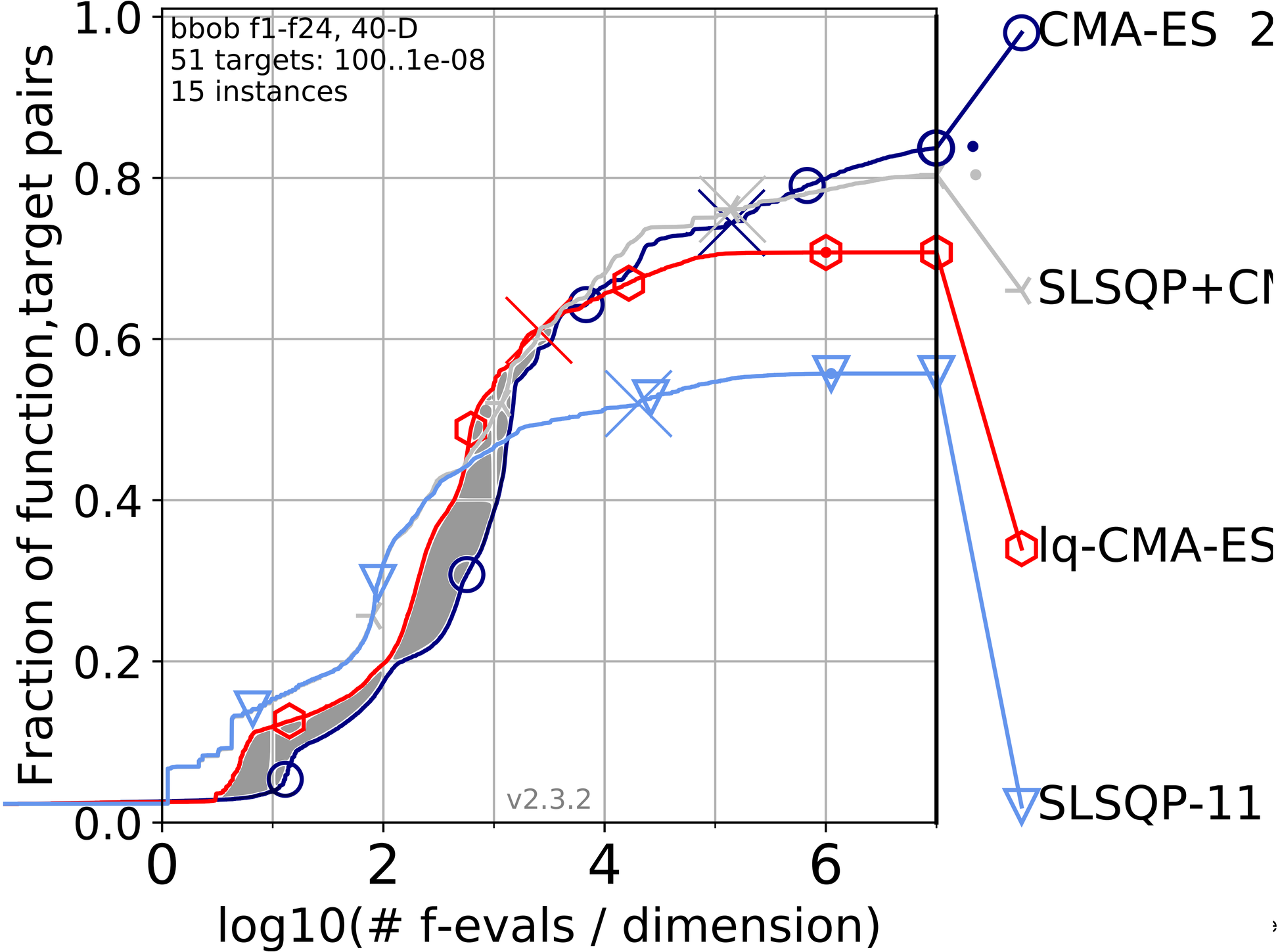
Aggregated Runtime Distributions

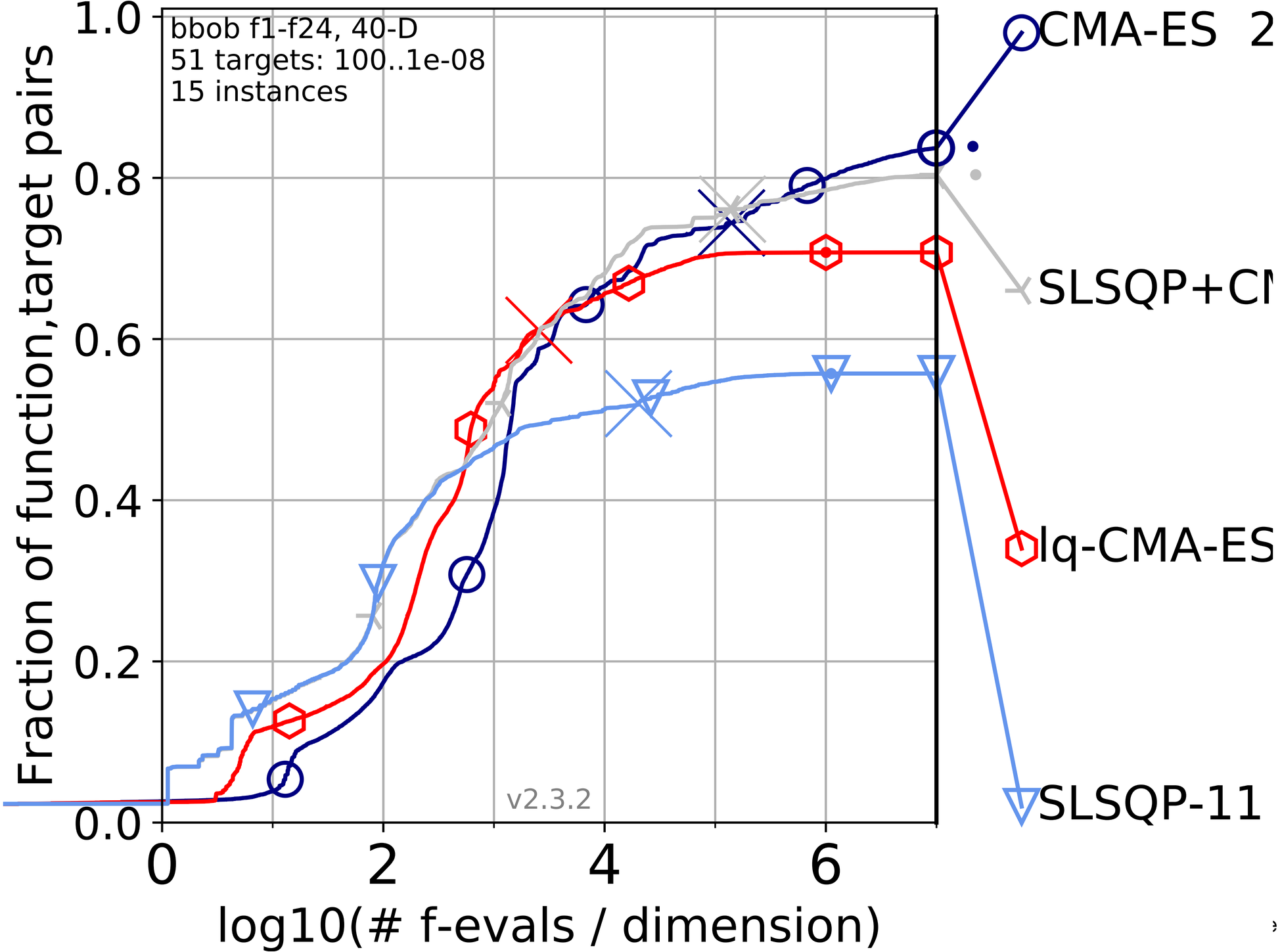




Hansen 2019. A Global Surrogate Assisted CMA-ES. *GECCO '19*.


bbob f1-f24, 40-D
51 targets: 100..1e-08
15 instances
v2.3.2
Fraction of function,target pairs
log10(# f-evals / dimension)
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
CMA-ES 2
SLSQP+CM
lq-CMA-ES
SLSQP-11


bbob f1-f24, 40-D
51 targets: 100..1e-08
15 instances
v2.3.2
Fraction of function,target pairs
log10(# f-evals / dimension)
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
CMA-ES 2
SLSQP+CM
lq-CMA-ES
SLSQP-11


bbob f1-f24, 40-D
51 targets: 100..1e-08
15 instances
v2.3.2
Fraction of function,target pairs
log10(# f-evals / dimension)
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
CMA-ES 2
SLSQP+CM
lq-CMA-ES
SLSQP-11

Recall: the *in*famous p-value

is the *probability of the observed (or more extreme) data to occur given the null hypothesis H_0 were 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 $p < 0.05$.

Recall: the infamous p-value

However,

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

and these may not even be close at all!

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

Recall: the infamous p-value

However,

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

and these may not even be close at all!

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

WRONG

(almost always)

(almost always)

- Gigerenzer 2018: "**editors should no longer accept manuscripts that report results as "significant" or "not significant.**" If *p* values are reported, they should be reported as exact *p* values"

  Statistical rituals: The replication delusion and how we got there. *Advances in Methods and Practices in Psychological Science 1.2 (2018): 198-218.*

- 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] 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).

- Gigerenzer 2018: "**editors should no longer accept manuscripts that report results as "significant" or "not significant.**" If *p* values are reported, they should be reported as exact *p* values"

  Statistical rituals: The replication delusion and how we got there. *Advances in Methods and Practices in Psychological Science 1.2 (2018): 198-218.*

- 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] 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).

- Gigerenzer 2018: "**editors should no longer accept manuscripts that report results as "significant" or "not significant.**" If *p* values are reported, they should be reported as exact *p* values"

  Statistical rituals: The replication delusion and how we got there. *Advances in Methods and Practices in Psychological Science 1.2 (2018): 198-218.*

- 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] 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).

- Gigerenzer 2018: "**editors should no longer accept manuscripts that report results as "significant" or "not significant.**" If *p* values are reported, they should be reported as exact *p* values"

  Statistical rituals: The replication delusion and how we got there. *Advances in Methods and Practices in Psychological Science 1.2 (2018): 198-218.*

- 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] 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 (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 the p-value properly?

From data to knowledge

The (famous) **Bayes' Rule** reads

$$P(H_0 \mid D)$$

From data to knowledge

The (famous) **Bayes' Rule** reads

$$P(H_0 \mid D) \quad = \quad P(H_0)$$

From data to knowledge

The (famous) **Bayes' Rule** reads

$$P(H_0 \mid D) \quad = \quad P(H_0) \quad \times \quad P(D \mid H_0)$$

From data to knowledge

The (famous) **Bayes' Rule** reads

$$P(H_0 \mid D) = \frac{P(H_0) \times P(D \mid H_0)}{P(D)}$$

proof…

From data to knowledge

The (famous) **Bayes' Rule** reads

$$P(D) \times P(H_0 \mid D) = P(H_0) \times P(D \mid H_0)$$

both sides equal to…

From data to knowledge

The (famous) **Bayes' Rule** reads

$$P(D) \times P(H_0 \mid D) = P(H_0) \times P(D \mid H_0)$$

Proof:

$$= P(H_0 \cap D)$$

From data to knowledge

The (famous) **Bayes' Rule** reads

$$P(D) \times P(H_0 \mid D) = P(H_0) \times P(D \mid H_0)$$

repeat for H1…

From data to knowledge

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

$$\frac{P(D) \times P(H_0 \mid D)}{P(D) \times P(H_1 \mid D)} = \frac{P(H_0)}{P(H_1)} \times \frac{P(D \mid H_0)}{P(D \mid H_1)}$$

posterior "odds" prior "odds" Bayes factor

Proof: Bayes' Theorem reads $P(H_0 \mid D) = P(H_0)\frac{P(D \mid H_0)}{P(D)}$ and likewise

$P(H_1 \mid D) = P(H_1)\frac{P(D \mid H_1)}{P(D)}$ 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

From data to knowledge

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

$$\underbrace{\frac{P(H_0 \mid D)}{P(H_1 \mid D)}}_{\text{posterior "odds"}} = \underbrace{\frac{P(H_0)}{P(H_1)}}_{\text{prior "odds"}} \times \underbrace{\frac{P(D \mid H_0)}{P(D \mid H_1)}}_{\text{Bayes factor}}$$

Proof: Bayes' Theorem reads $P(H_0 \mid D) = P(H_0)\frac{P(D \mid H_0)}{P(D)}$ and likewise $P(H_1 \mid D) = P(H_1)\frac{P(D \mid H_1)}{P(D)}$ 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

From data to knowledge

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

$$\underbrace{\frac{P(H_0 \mid D)}{P(H_1 \mid D)}}_{\text{posterior "odds"}} = \underbrace{\frac{P(H_0)}{P(H_1)}}_{\text{prior "odds"}} \times \underbrace{\frac{P(D \mid H_0)}{P(D \mid H_1)}}_{\text{Bayes factor}}$$

Proof: Bayes' Theorem reads $P(H_0 \mid D) = P(H_0)\frac{P(D \mid H_0)}{P(D)}$ and likewise $P(H_1 \mid D) = P(H_1)\frac{P(D \mid H_1)}{P(D)}$ and we divide the two equations.

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

From data to knowledge

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

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

Proof: Bayes' Theorem reads $P(H_0 \mid D) = P(H_0)\frac{P(D \mid H_0)}{P(D)}$ and likewise $P(H_1 \mid D) = P(H_1)\frac{P(D \mid H_1)}{P(D)}$ and we divide the two equations.

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

From data to knowledge

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

If $H_1 = \neg H_0$, 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

From data to knowledge

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

If $H_1 = \neg H_0$, 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

From data to knowledge

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

If $H_1 = \neg H_0$, the (famous) **Bayes' Rule in "odds form"** reads

$$\underbrace{O(H_0 \mid D)}_{\text{posterior odds}} = \overbrace{\underbrace{O(H_0)}_{\text{prior odds}}}^{1000:1} \times \underbrace{\frac{P(D \mid H_0)}{P(D \mid \neg H_0)}}_{\text{Bayes factor}}$$

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

From data to knowledge

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

If $H_1 = \neg H_0$, the (famous) **Bayes' Rule in "odds form"** reads

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

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

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

If $H_1 = \neg H_0$, the (famous) **Bayes' Rule in "odds form"** reads

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

$10 : 1$ $\qquad$ $1000 : 1$ $\qquad$ $\frac{1/100}{1}$

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 $P(D \mid H_0) \approx p$

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

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 $P(D \mid \neg H_0) \approx \exp(-1)$

i.e., D is not untypical as observation when $\neg H_0$ is true, $E \log(U[0,1]) = -1$

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 $P(D \mid \neg H_0) \approx \exp(-1)$

i.e., D is not untypical as observation when $\neg H_0$ is true, $E \log(U[0,1]) = -1$

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

$$\underbrace{O(H_0 \mid D)}_{\text{posterior odds}} \approx \underbrace{O(H_0)}_{\text{prior odds}} \times 2.7\,p$$

as a rough approximation for the posterior odds of H_0.

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

$$\underbrace{O(H_0 \mid D)}_{\text{posterior odds}} \approx \underbrace{O(H_0)}_{\text{prior odds}} \times 2.7\,p$$

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.

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

$$\underbrace{O(H_0 \mid D)}_{\text{posterior odds}} \approx \underbrace{O(H_0)}_{\text{prior odds}} \times 2.7\,p$$

as a rough approximation for the posterior odds of H_0.

Extraordinary claims require extraordinary evidence.

Sagan 1979. Broca's Brain.

To claim that the Odds$(H_0 \mid D)$ is small given the Odds$(H_0) \gg 1$ is extraordinary and **requires** $p \ll 1 \,/\, \text{Odds}(H_0)$.

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!

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

$$\approx \underbrace{O(H_0)}_{\text{prior odds}} \times \prod_{i=1}^{k} 2.7\, p_i$$

When the $p_i \ll 1/3$, the probability of H_0 vanishes quickly (geometrically fast) with the number of independent replications.

From data to knowledge: Summary

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

From data to knowledge: Summary

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

$$\underbrace{\frac{P(H_0 \mid D)}{P(\neg H_0 \mid D)}}_{\text{posterior "odds"}} = \underbrace{\frac{P(H_0)}{P(\neg H_0)}}_{\text{prior "odds"}} \times \underbrace{\frac{P(D \mid H_0)}{P(D \mid \neg H_0)}}_{\text{Bayes factor}} \approx \underbrace{\frac{P(H_0)}{P(\neg H_0)}}_{\text{prior "odds"}} \times 3\,p$$

From data to knowledge: Summary

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

$$\underbrace{\frac{P(H_0 \mid D)}{P(\neg H_0 \mid D)}}_{\text{posterior "odds"}} = \underbrace{\frac{P(H_0)}{P(\neg H_0)}}_{\text{prior "odds"}} \times \underbrace{\frac{P(D \mid H_0)}{P(D \mid \neg H_0)}}_{\text{Bayes factor}} \approx \underbrace{\frac{P(H_0)}{P(\neg H_0)}}_{\text{prior "odds"}} \times 3\,p$$

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

a small p stands on its own merits: we *can* conclude that the odds for H_0 have decreased by a factor of about $3p$

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, 10^4. This requires $p = 10^{-6}$ to reject H_0 with an error probability of ≈ 0.03, assuming that the experiment was conducted without any flaws! If we additionally account for the probability of (honest or dishonest) mistakes, $p < 10^{-4}$ is rather hard to achieve and we need independent confirmation.

the higher the prior odds for H_0, the more exceptional and surprising is the scientific result and the smaller needs to be p in order to reject H_0 with some confidence

From data to knowledge: Summary

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

$$\underbrace{\frac{P(H_0 \mid D)}{P(\neg H_0 \mid D)}}_{\text{posterior "odds"}} \approx \underbrace{\frac{P(H_0)}{P(\neg H_0)}}_{\text{prior "odds"}} \times 3p$$

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

a small p stands on its own merits: we *can* conclude that the odds for H_0 have decreased by a factor of about $3p$

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, 10^4. This requires $p = 10^{-6}$ to reject H_0 with an error probability of ≈ 0.03, assuming that the experiment was conducted without any flaws! If we additionally account for the probability of (honest or dishonest) mistakes, $p < 10^{-4}$ is rather hard to achieve and we need independent confirmation.

the higher the prior odds for H_0, the more exceptional and surprising is the scientific result and the smaller needs to be p in order to reject H_0 with some confidence

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 *changes* **by a factor** of about $2.7p$ (namely, it decreases when $p < \exp(-1)$)

not: ~~$p < 0.05$ implies that the odds to mistakenly reject H_0 are small~~

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 reliable is the papers methodology? How often has the paper **been replicated**? What is the current (posterior) odds for H_0?
    $\Longrightarrow$ quantification of the confidence in replicability.

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)