AutoML: Evaluation

Visualizing Evaluation over Time

Bernd Bischl Frank Hutter Lars Kotthoff
Marius Lindauer

Motivation

- If we define AutoML as an optimization process, the incumbent solution (i.e., the best found configuration so far) gradually improves over time

Motivation

- If we define AutoML as an optimization process, the incumbent solution (i.e., the best found configuration so far) gradually improves over time
- We don't know when users will stop the AutoML process
  - Running over the coffee break (15min)
  - Running during a meeting (1h)
  - Running over night (16h)
  - Running over the weekend (48+h)

Motivation

- If we define AutoML as an optimization process, the incumbent solution (i.e., the best found configuration so far) gradually improves over time
- We don't know when users will stop the AutoML process
  - Running over the coffee break (15min)
  - Running during a meeting (1h)
  - Running over night (16h)
  - Running over the weekend (48+h)

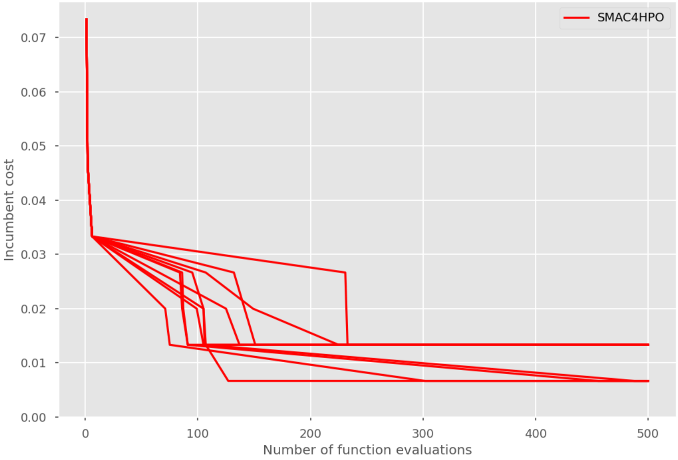
$\rightsquigarrow$ Anytime performance of AutoML is important

  - i.e., the AutoML tool should return the best possible solution at each time point

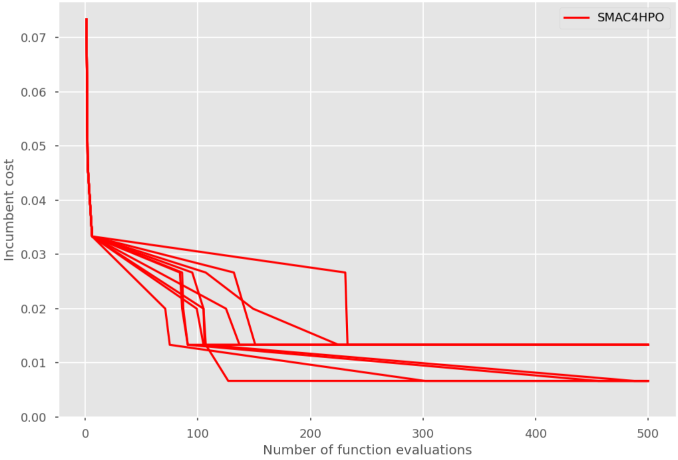
Observing Performance over Time

(*Empty slides for drawing something live in the video.*)

Repeated Experiments

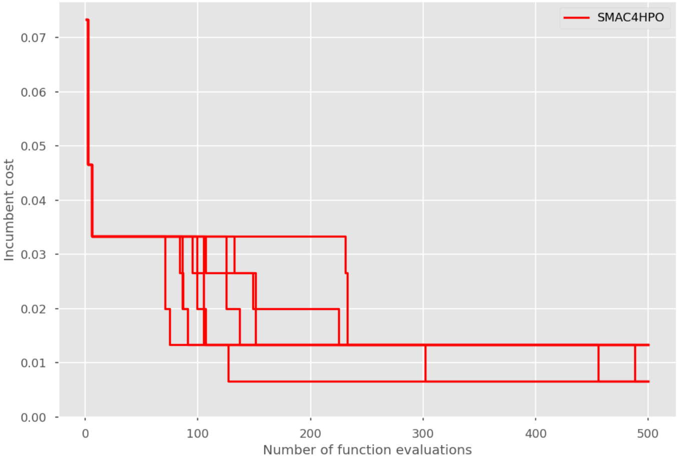


Repeated Experiments

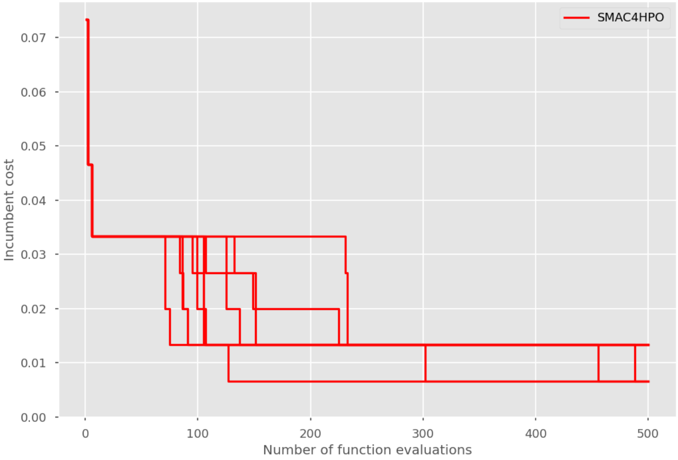


$\rightsquigarrow$ Don't linearly interpolate between points!

Step Functions

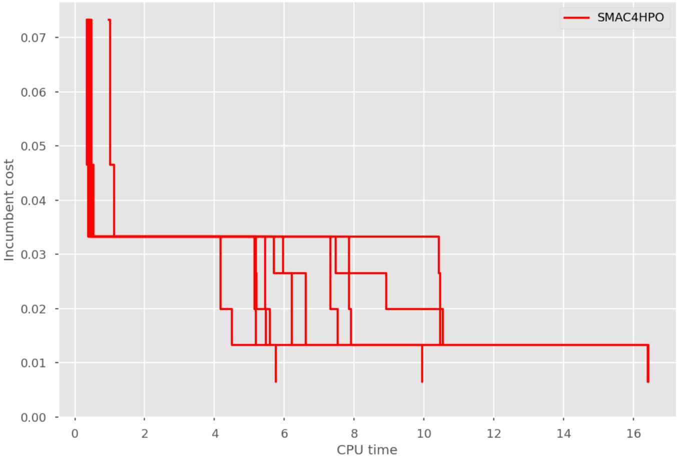


Step Functions



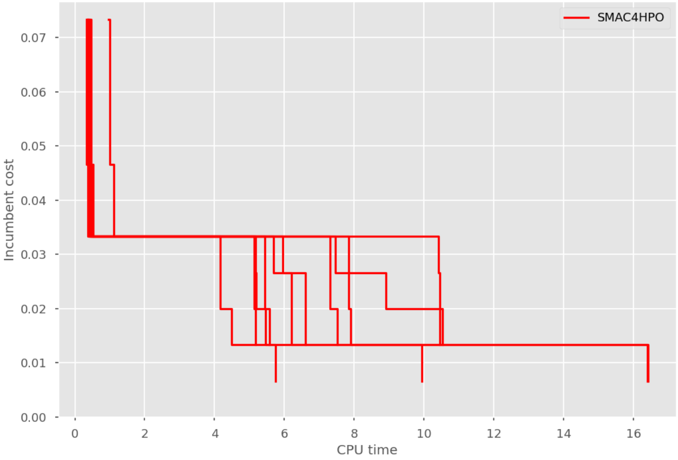
$\rightsquigarrow$ Do we care about number of function evaluations?

CPU Time



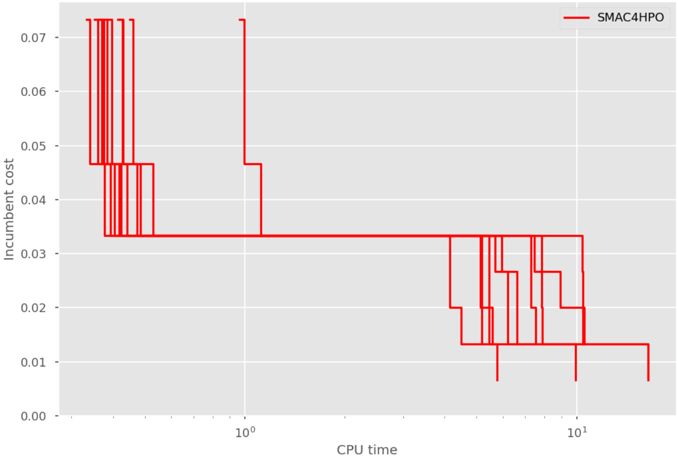
SMAC4HPO
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
Incumbent cost
0
2
4
6
8
10
12
14
16
CPU time

CPU Time



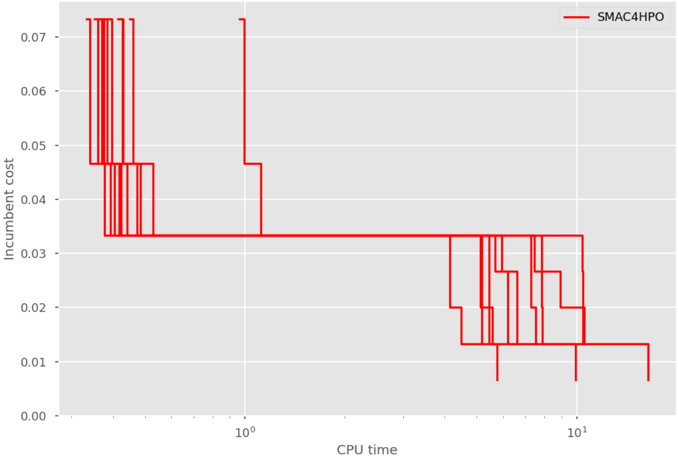
$\rightsquigarrow$ We might loose information in the beginning.

x-log scale



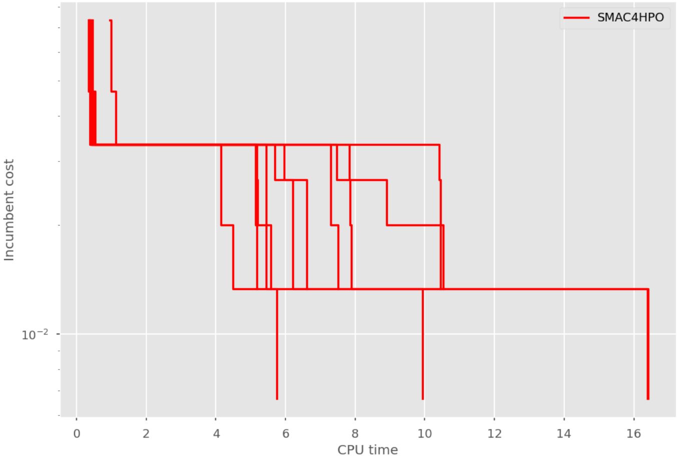
SMAC4HPO
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
Incumbent cost
10^0
10^1
CPU time

x-log scale

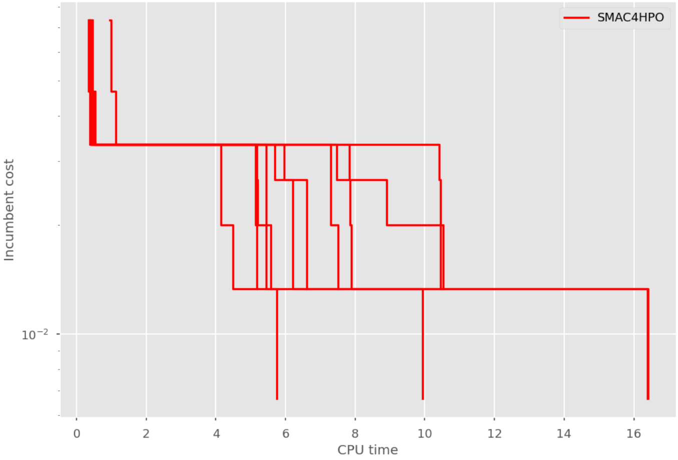


$\leadsto$ Small differences on y are hard to spot.

y-log scale

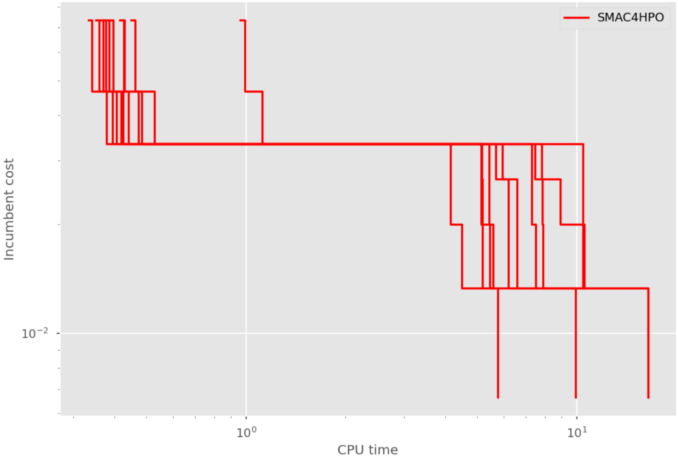


y-log scale

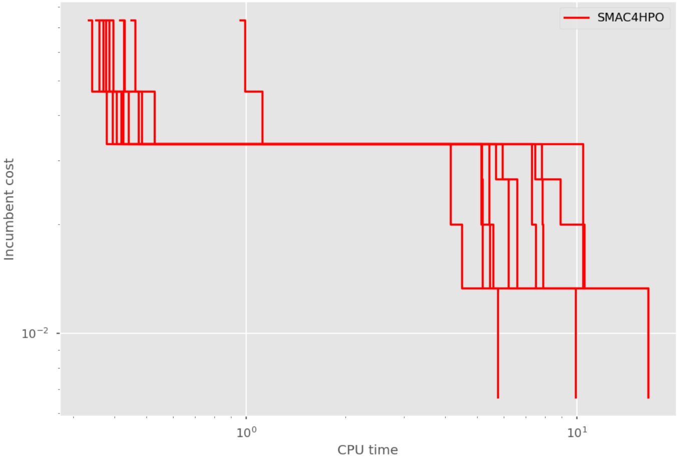


$\rightsquigarrow$ Log on both?

x-y-log scale

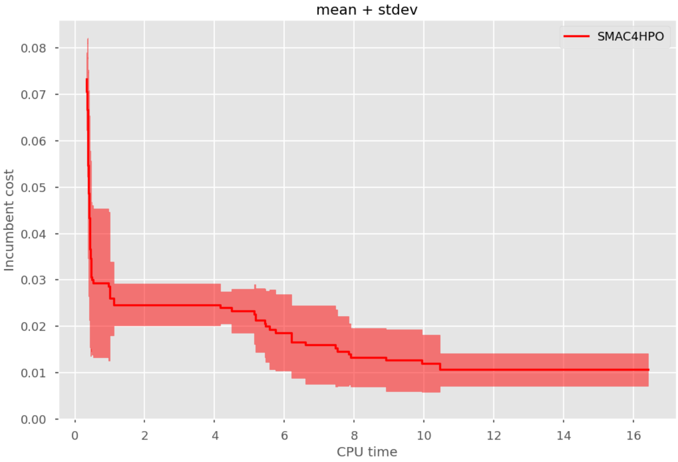


x-y-log scale

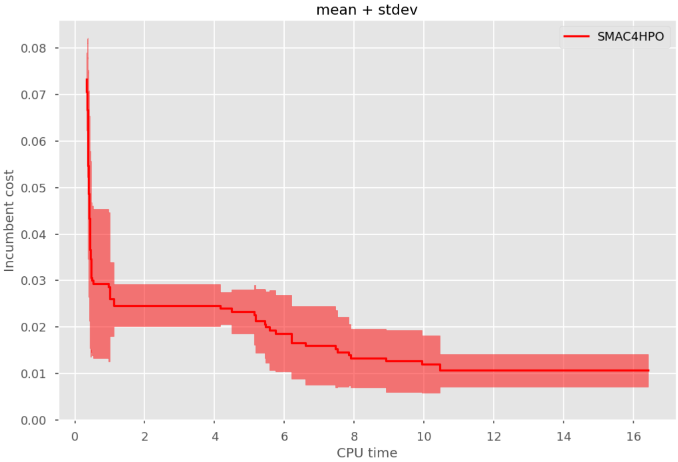


$\rightsquigarrow$ Can we summarize the individual curves?

Mean ± Standard Deviation: $\mu \pm \sigma$

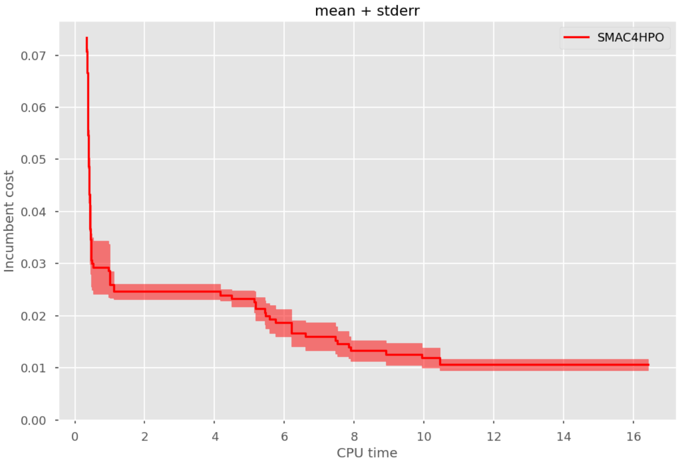


Mean $\pm$ Standard Deviation: $\mu \pm \sigma$

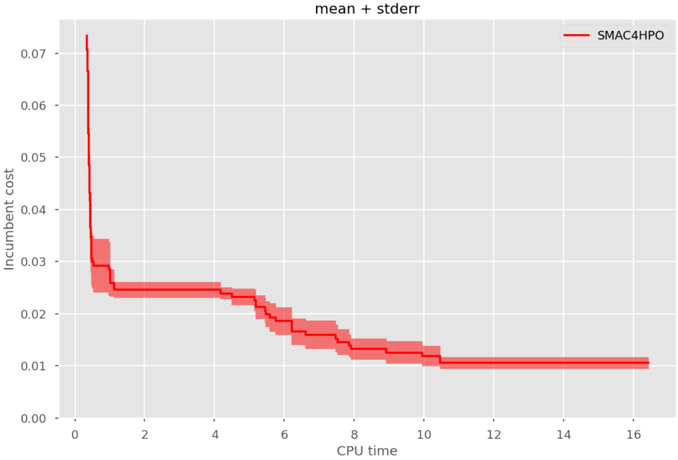


$\rightsquigarrow$ Mean $\pm$ standard deviation works only if uncertainty is symmetric.

Mean $\pm$ Standard Error: $\mu \pm \frac{\sigma}{\sqrt{n}}$

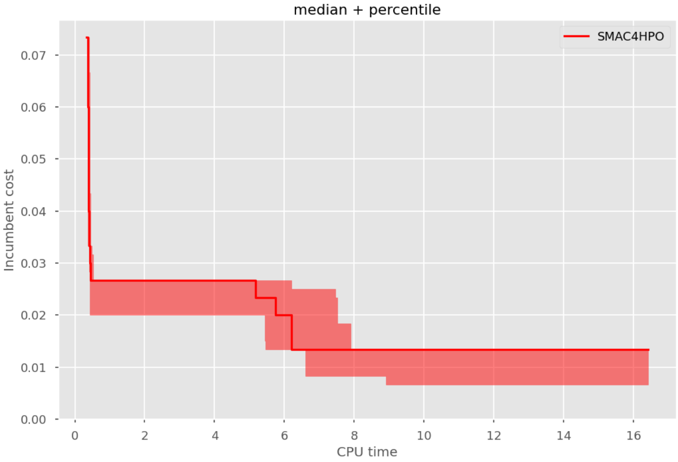


Mean $\pm$ Standard Error: $\mu \pm \frac{\sigma}{\sqrt{n}}$

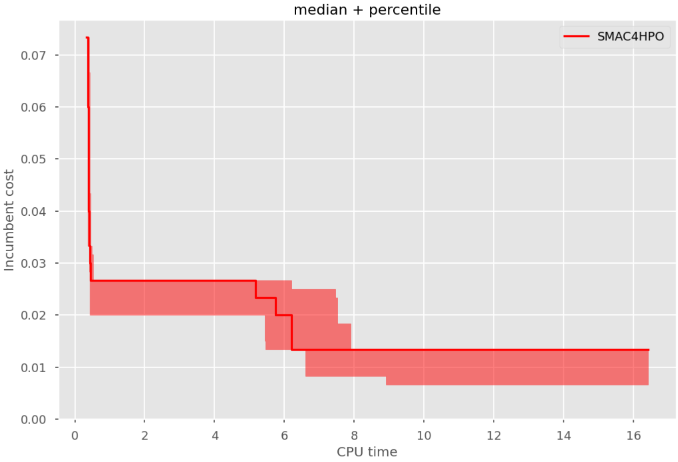


$\rightsquigarrow$ Confidence of the mean estimate!

Median + 25/75th Percentile

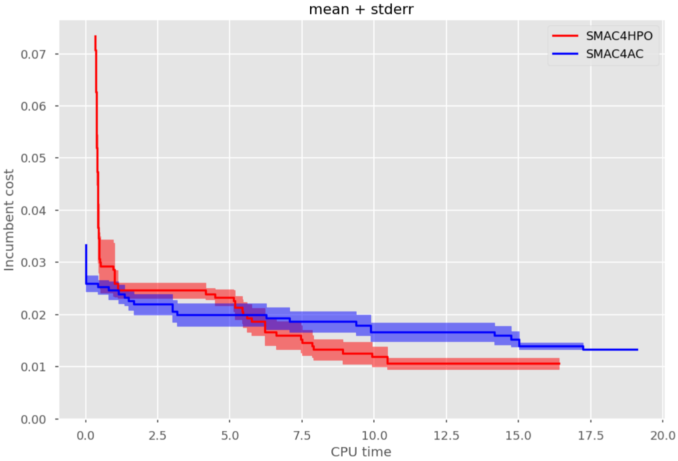


Median + 25/75th Percentile



$\rightsquigarrow$ Works also for asymmetric uncertainties.

Comparing 2 AutoML Optimizers



Summary

1. Plotting anytime performance is helpful
2. On real benchmarks often better to plot CPU time instead of function evaluations
3. Use step functions!
4. Consider log-scales on x and/or y
5. Consider different ways for plotting the uncertainty of cost observations