Intelligence.
Beautifully engineered.

No more grid search! How to build models effectively

Thomas Huijskens

London Driverless AI meet-up
1st August, 2019

QUANTUMBLACK
A MCKINSEY COMPANY

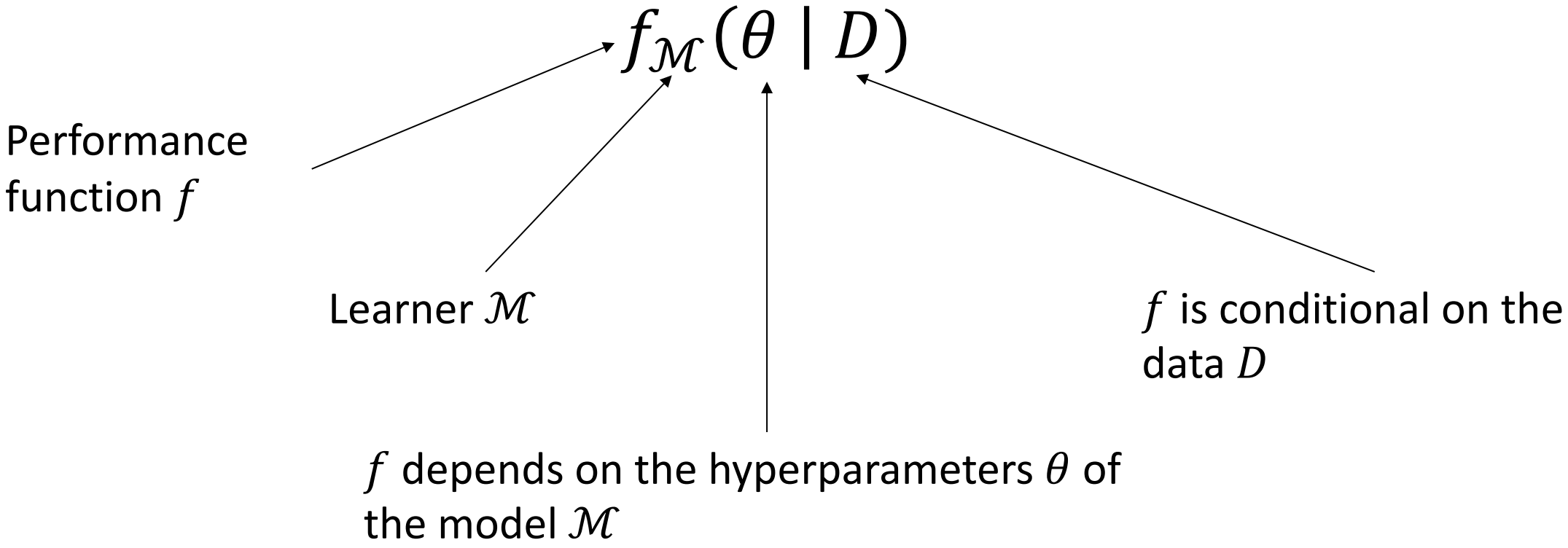
Confidential and proprietary: Any use of this material without
specific permission of McKinsey & Company is strictly prohibited

Most models have a number of hyperparameters that need to be chosen a priori..

- Typically, we want to optimize some performance metric $f_{\mathcal{M}}$ for a model $\mathcal{M}$, on a hold-out set of data.
- The model $\mathcal{M}$ has some hyperparameters θ that we need to specify, and the performance of $\mathcal{M}$ is highly dependent on the settings of θ.
- Example: For a classification problem, $\mathcal{M}$ could be a support vector machine, and the performance metric $f_{\mathcal{M}}$ could be the out-of-sample AUC.
- Because the performance of the SVM depends on hyperparameters θ, the metric $f_{\mathcal{M}}$ depends on θ as well.

All content copyright © 2019 QuantumBlack, a McKinsey company

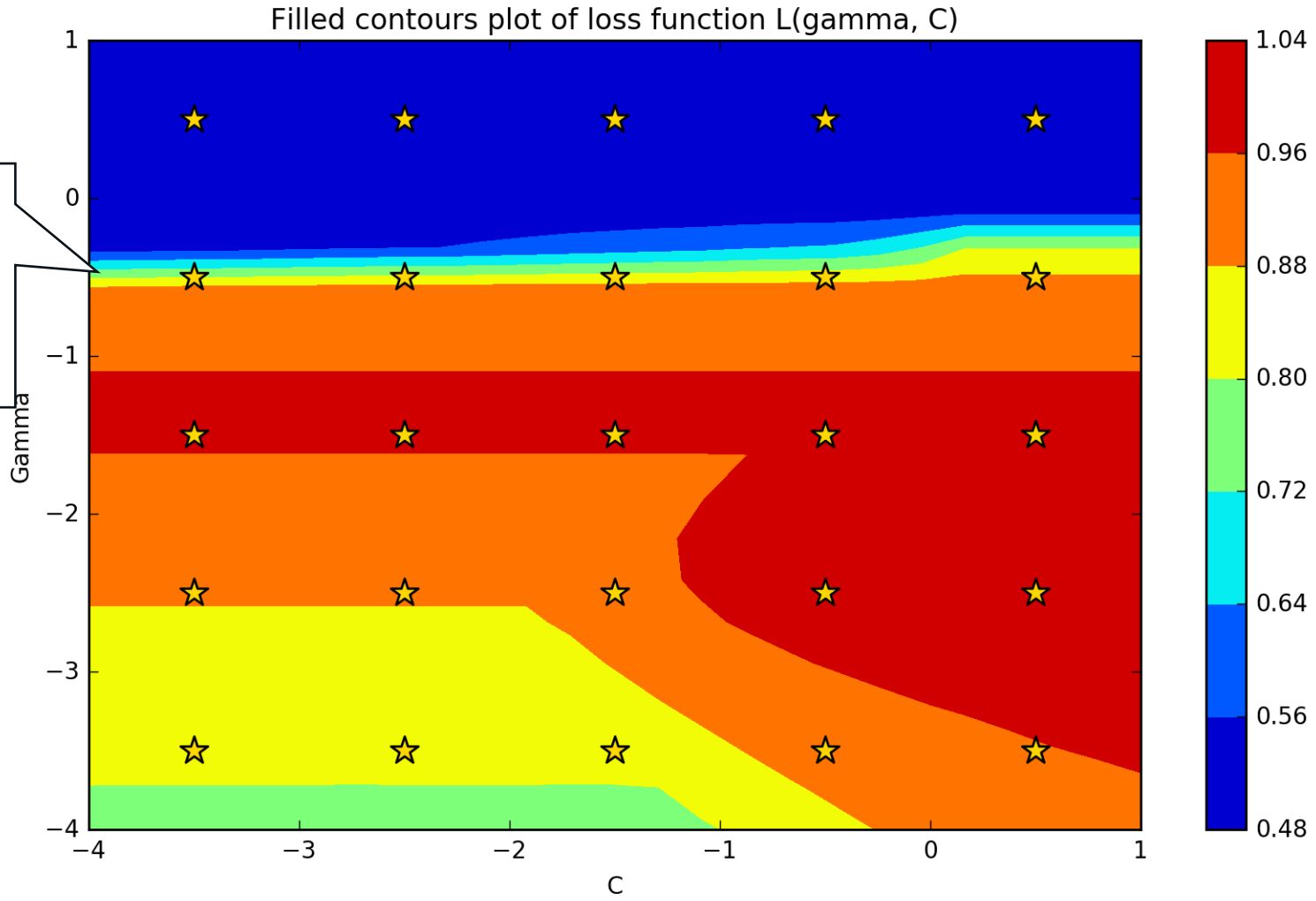
.. and our goal is to find the values of θ for which the value of the (out-of-sample) performance $f_{\mathcal{M}}$ is optimal



All content copyright © 2019 QuantumBlack, a McKinsey company

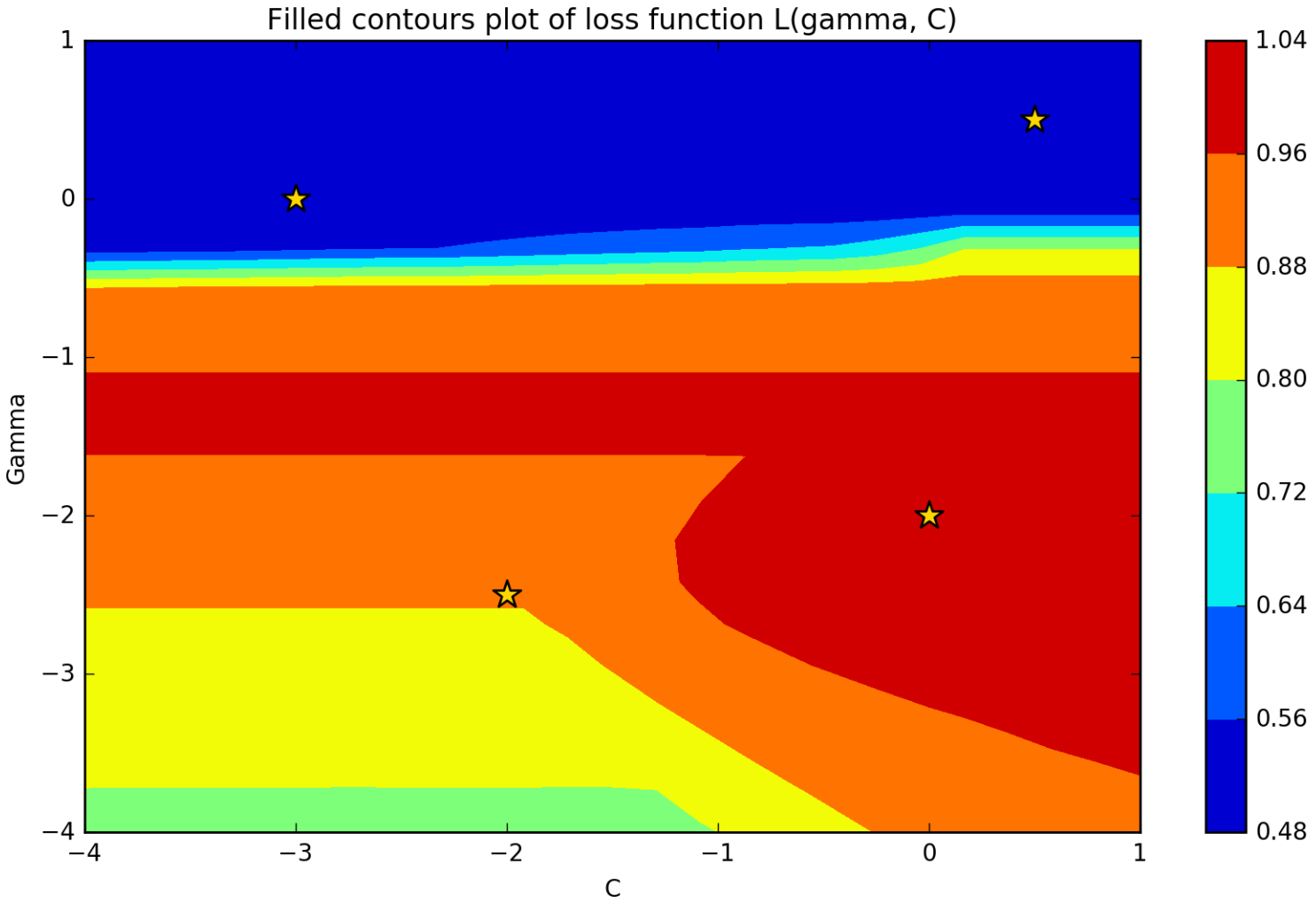
Manual grid search works for low-dimensional problems, but does not scale well to higher dimensions

Curse of dimensionality makes this inefficient in higher dimensional problems



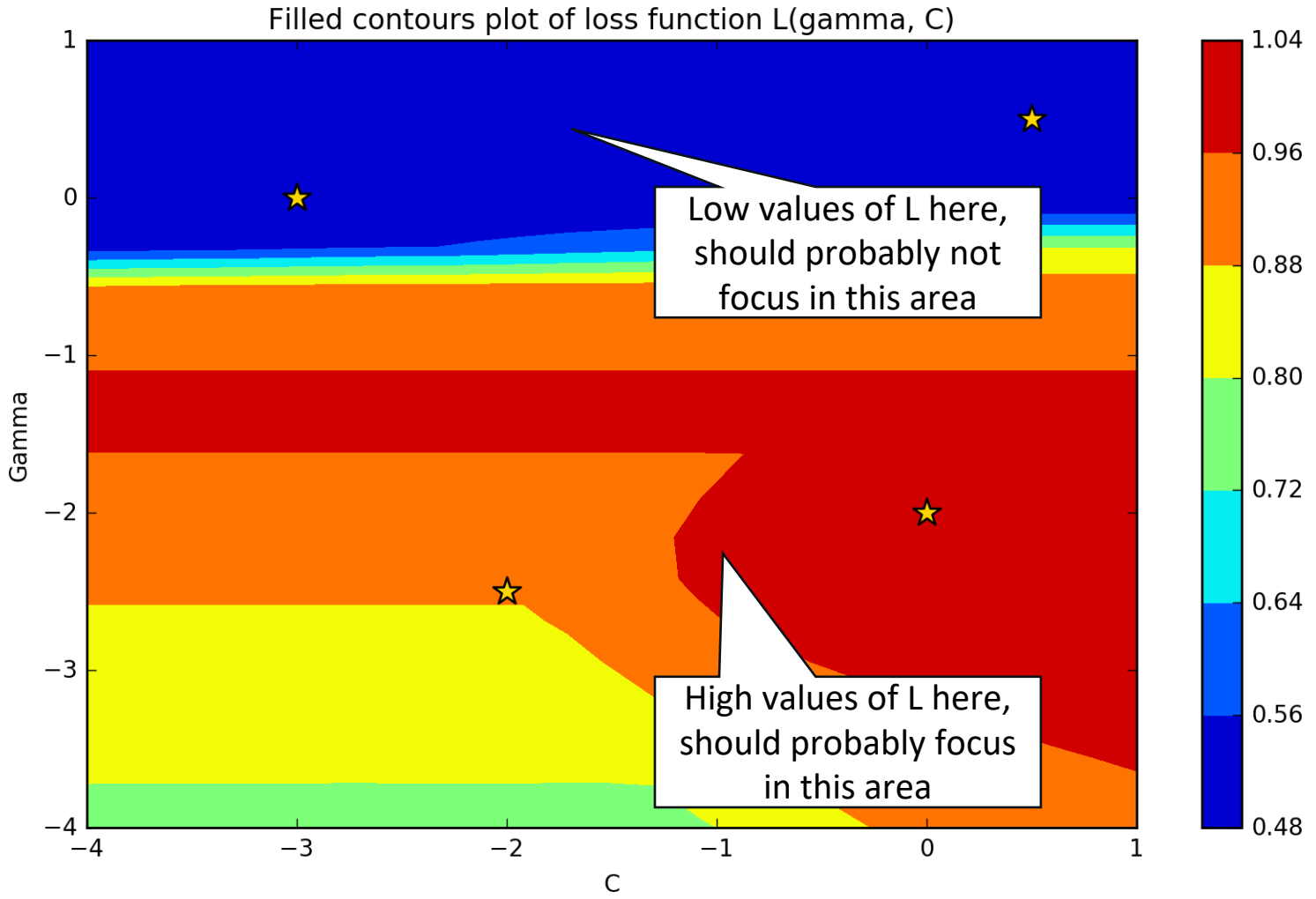
All content copyright © 2019 QuantumBlack, a McKinsey company

Random grid search works better in higher dimensions, but..



All content copyright © 2019 QuantumBlack, a McKinsey company

.. shouldn't previously evaluated hyperparameter values guide us in the search for the true optimal value?



All content copyright © 2019 QuantumBlack, a McKinsey company

Advanced hyper-parameter optimization methods

All content copyright © 2019 QuantumBlack, a McKinsey company

Advanced HPO methods are only beneficial in certain situations

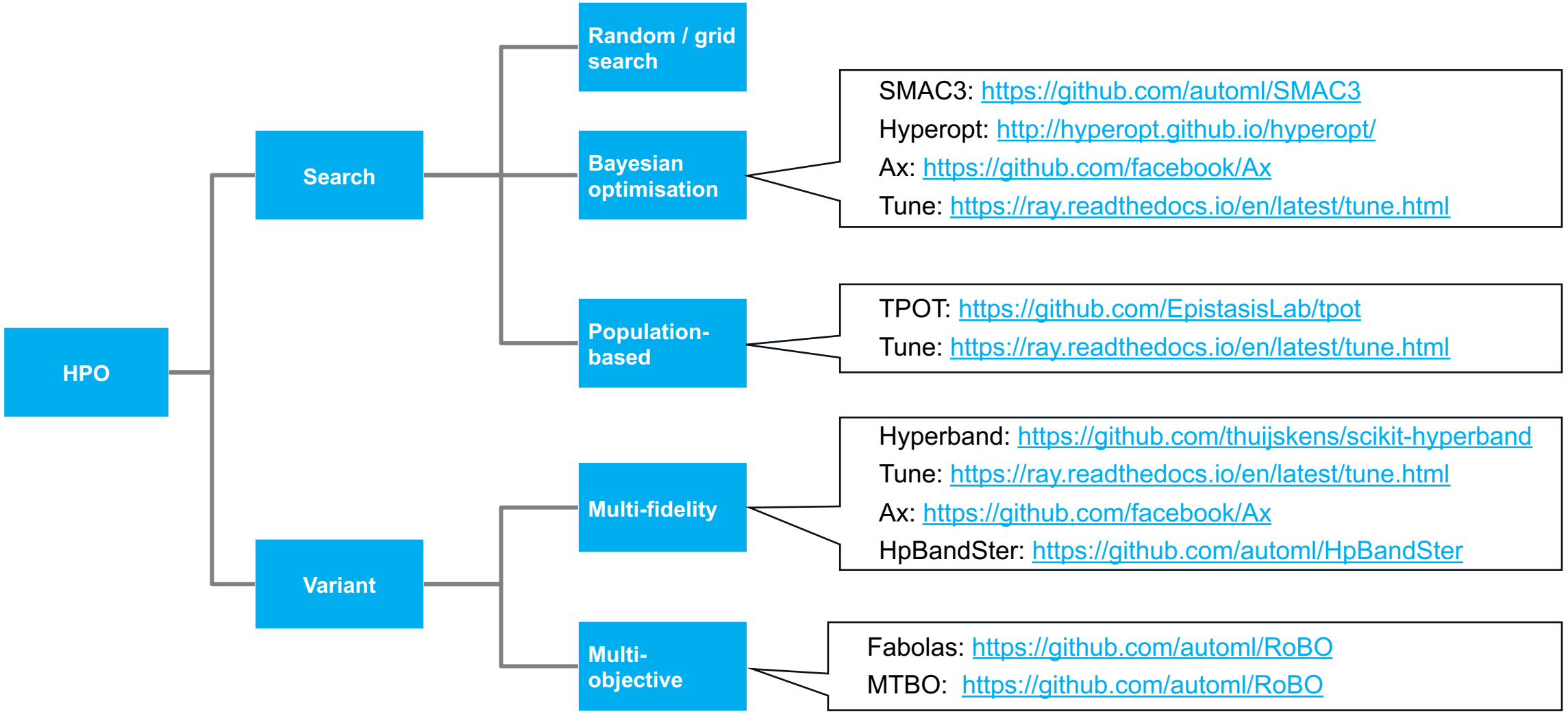
In advanced optimisation methods, we spend more computational time to figure out what set of parameters θ we are evaluating the performance metric $f_{\mathcal{M}}(\theta)$ for.

As a result, applying these methods only really makes sense if

- The number of hyperparameters is very high (θ is of high dimension); or
- It is computationally very expensive to evaluate $f_{\mathcal{M}}(\theta)$ for a single point θ.

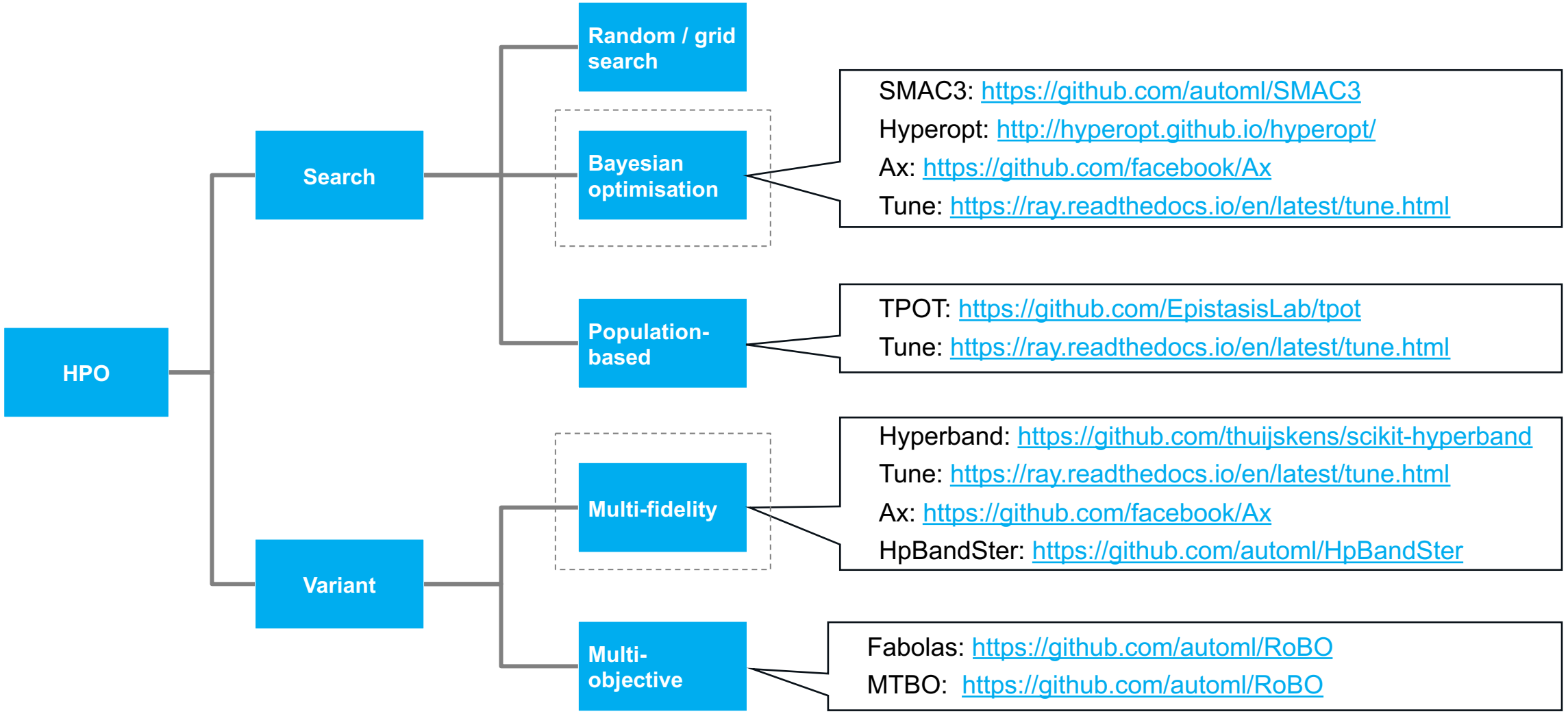
All content copyright © 2019 QuantumBlack, a McKinsey company

There are now a wide range of hyper-parameter optimisation methods



All content copyright © 2019 QuantumBlack, a McKinsey company

BO-HB is a recent method that combines Bayesian optimisation and multi-fidelity methods

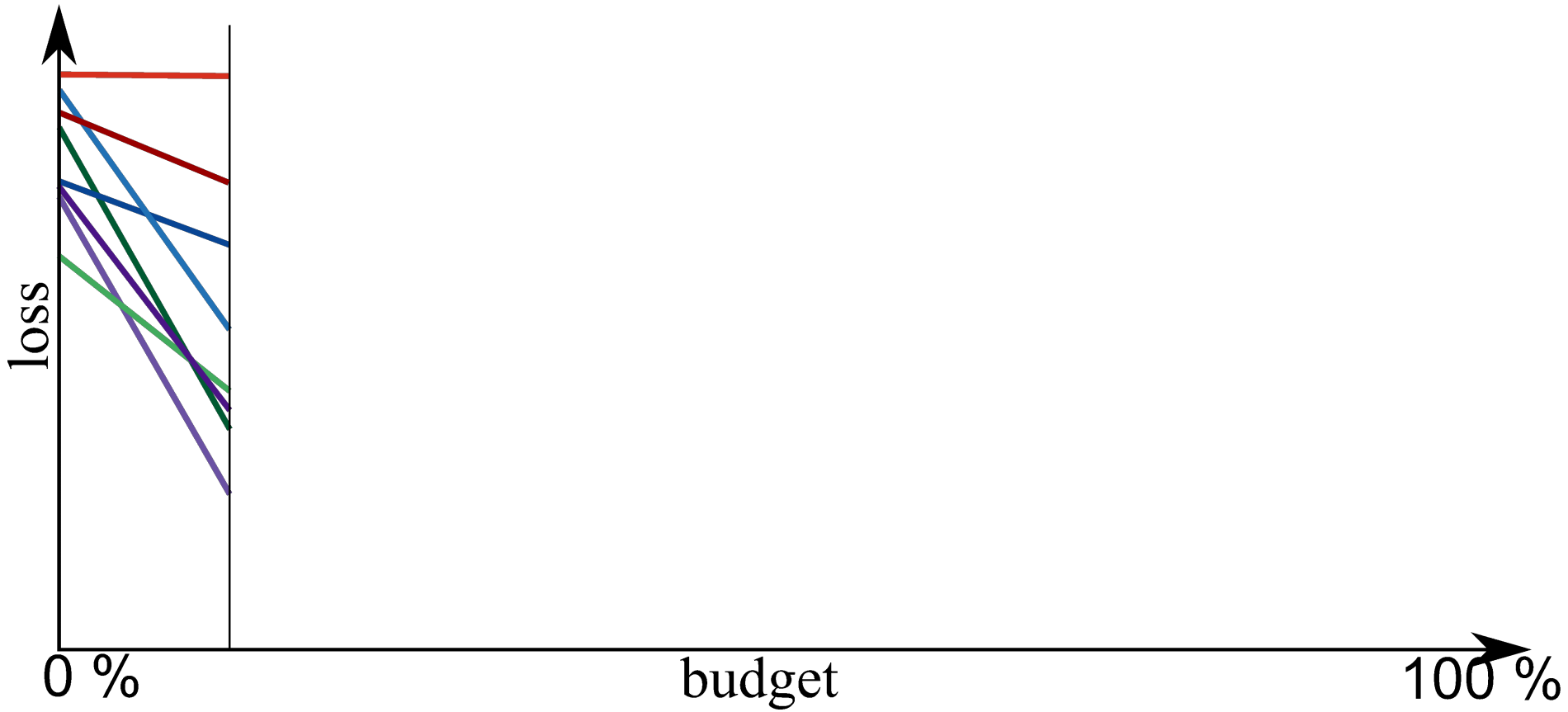


All content copyright © 2019 QuantumBlack, a McKinsey company

Hyperband

All content copyright © 2019 QuantumBlack, a McKinsey company

Hyperband iteratively runs the *successive halving* subroutine on model configurations



BOHB: Robust and efficient hyperparameter optimization at scale: https://www.automl.org/blog_bohb/

All content copyright © 2019 QuantumBlack, a McKinsey company

Successive halving suffers from the exploitation versus exploration trade-off

Given a total computational budget, the user has to decide whether

- To try many configurations with a small budget each; or
- To try only a few configurations with a larger budget.

Hyperband hedges for this trade-off by:

- Dividing the total computational budget into several **combinations** of number of configurations vs. budget.
- Call successive halving as a subroutine for each combination.

All content copyright © 2019 QuantumBlack, a McKinsey company

Number of configurations

Computational budget

Successive halving iteration

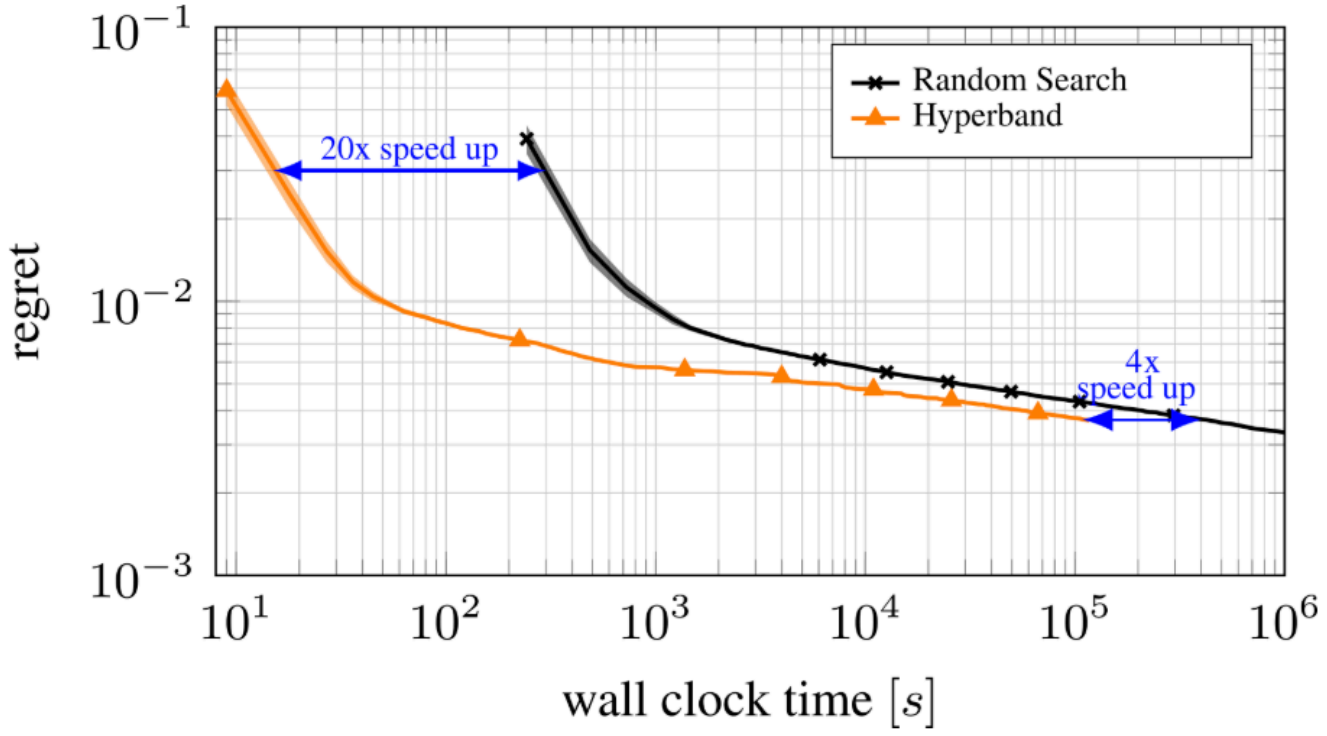
	$s=4$		$s=3$		$s=2$		$s=1$		$s=0$	
i	n_i	r_i	n_i	r_i	n_i	r_i	n_i	r_i	n_i	r_i
0	81	1	27	3	9	9	6	27	5	81
1	27	3	9	9	3	27	2	81		
2	9	9	3	27	1	81				
3	3	27	1	81						
4	1	81								

Table 1: The values of n_i and r_i for the brackets of HYPERBAND corresponding to various values of s, when $R = 81$ and $\eta = 3$.

Hyperband: A Novel Bandit-Based Approach to Hyperparameter Optimization, *Lisha Li, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, Ameet Talwalkar*, 2018. https://arxiv.org/abs/1603.06560

All content copyright © 2019 QuantumBlack, a McKinsey company

Hyperband's hedging strategy is very competitive against random search



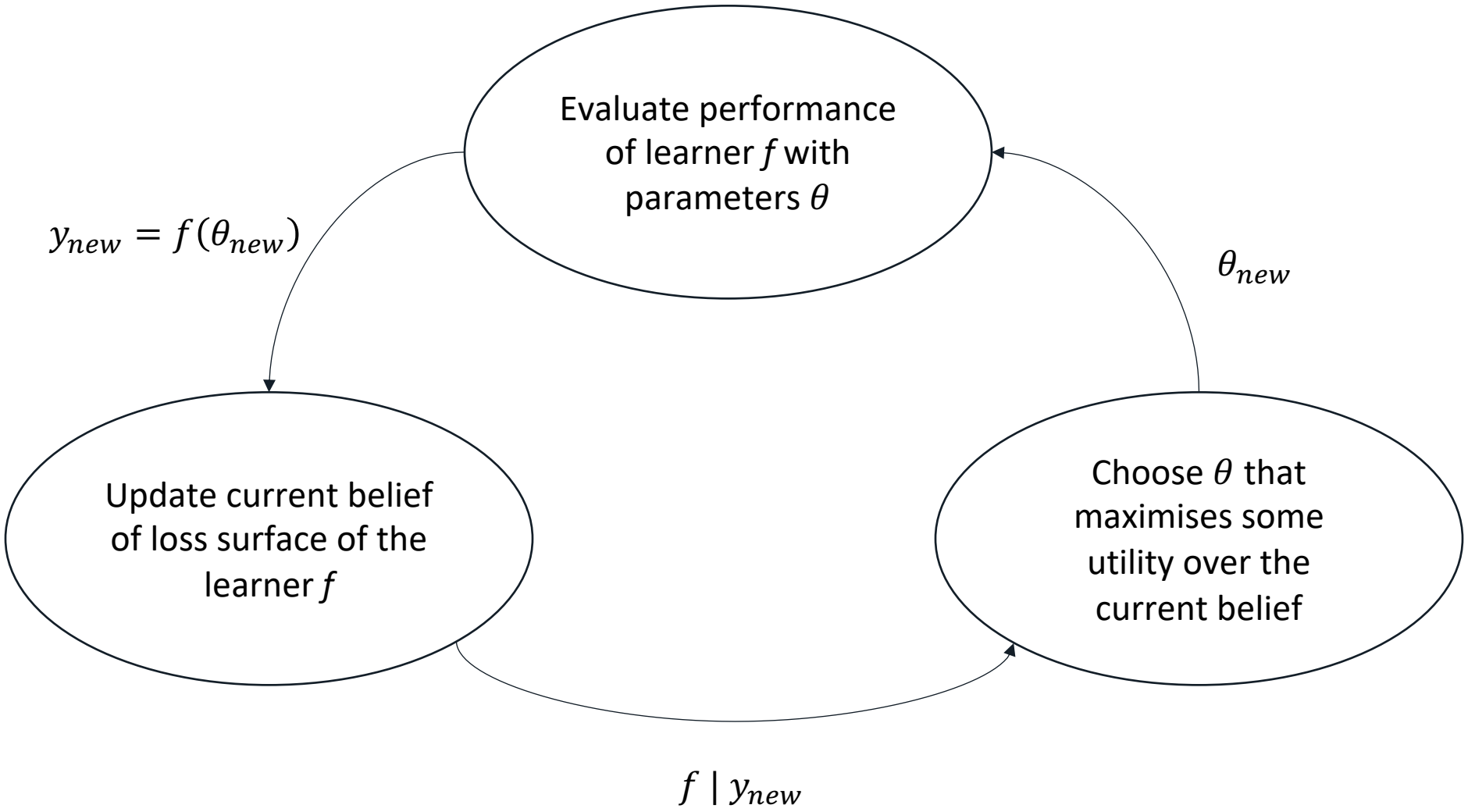
BOHB: Robust and Efficient Hyperparameter Optimization at Scale, ***Stefan Falkner, Aaron Klein, Frank Hutter*****, 2018.** https://arxiv.org/abs/1807.01774

All content copyright © 2019 QuantumBlack, a McKinsey company

Bayesian optimisation

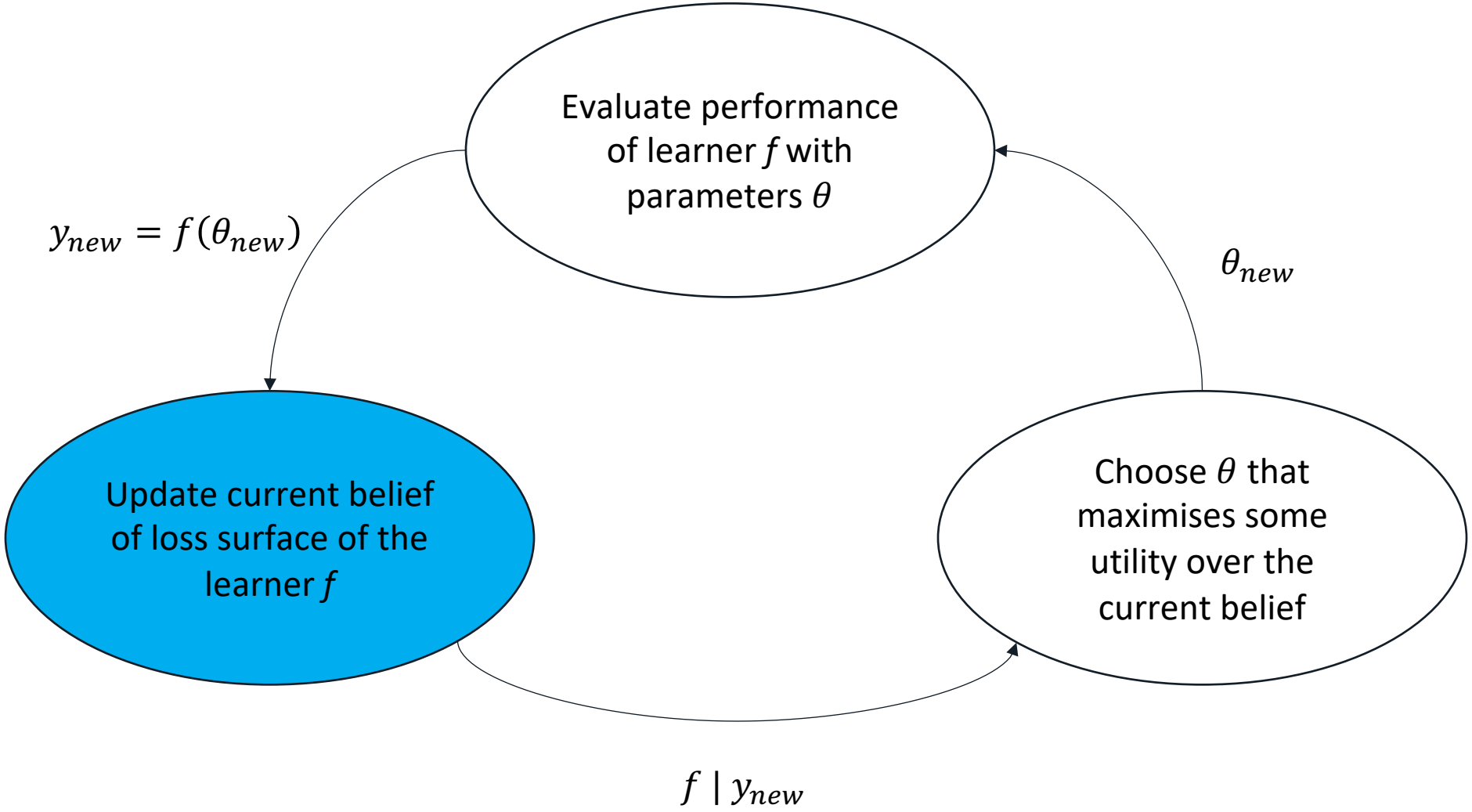
All content copyright © 2019 QuantumBlack, a McKinsey company

The classic Bayesian optimization algorithm consists of three steps



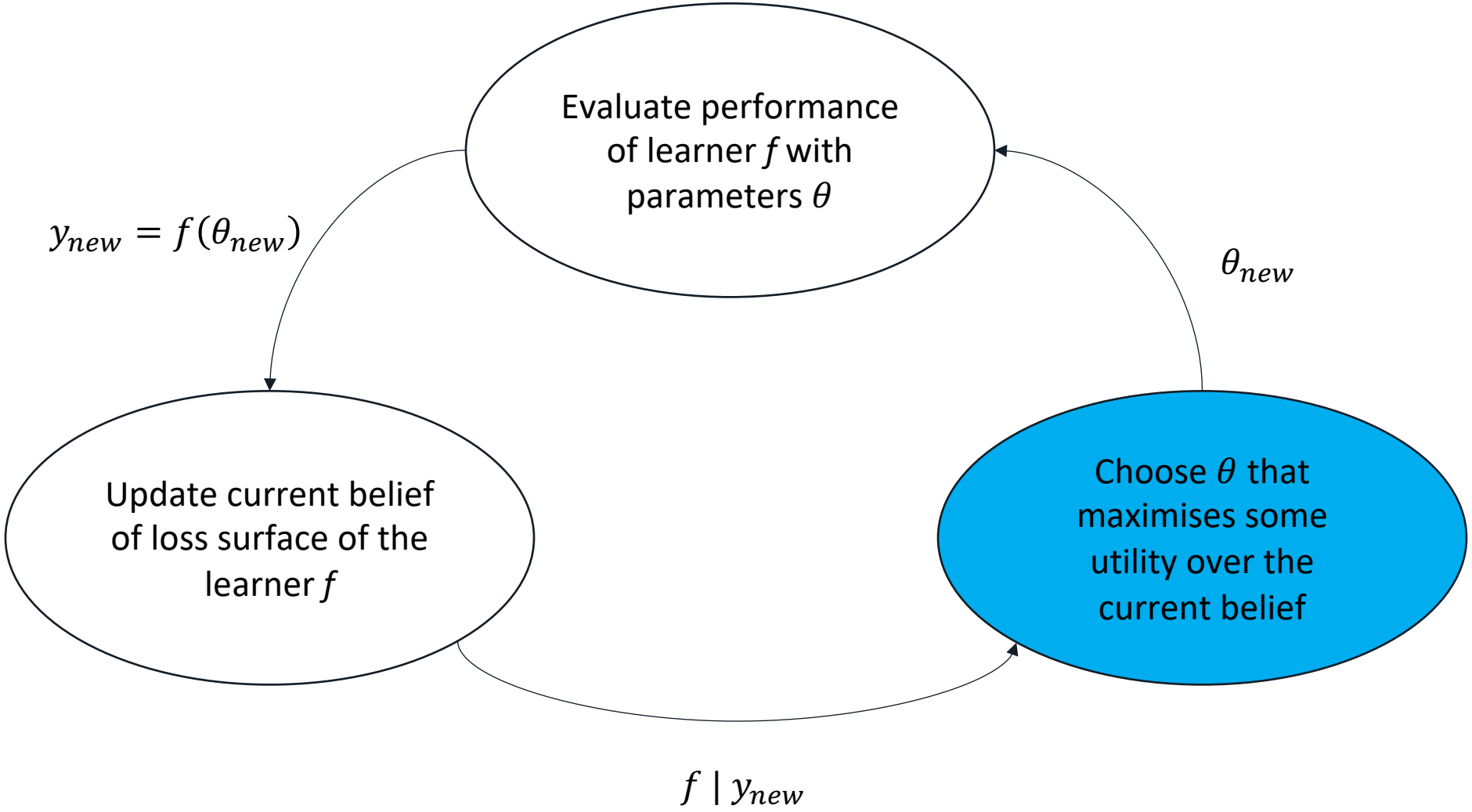
All content copyright © 2019 QuantumBlack, a McKinsey company

The performance function $f_{\mathcal{M}}$ is modelled to form a current belief of what the loss surface looks like..



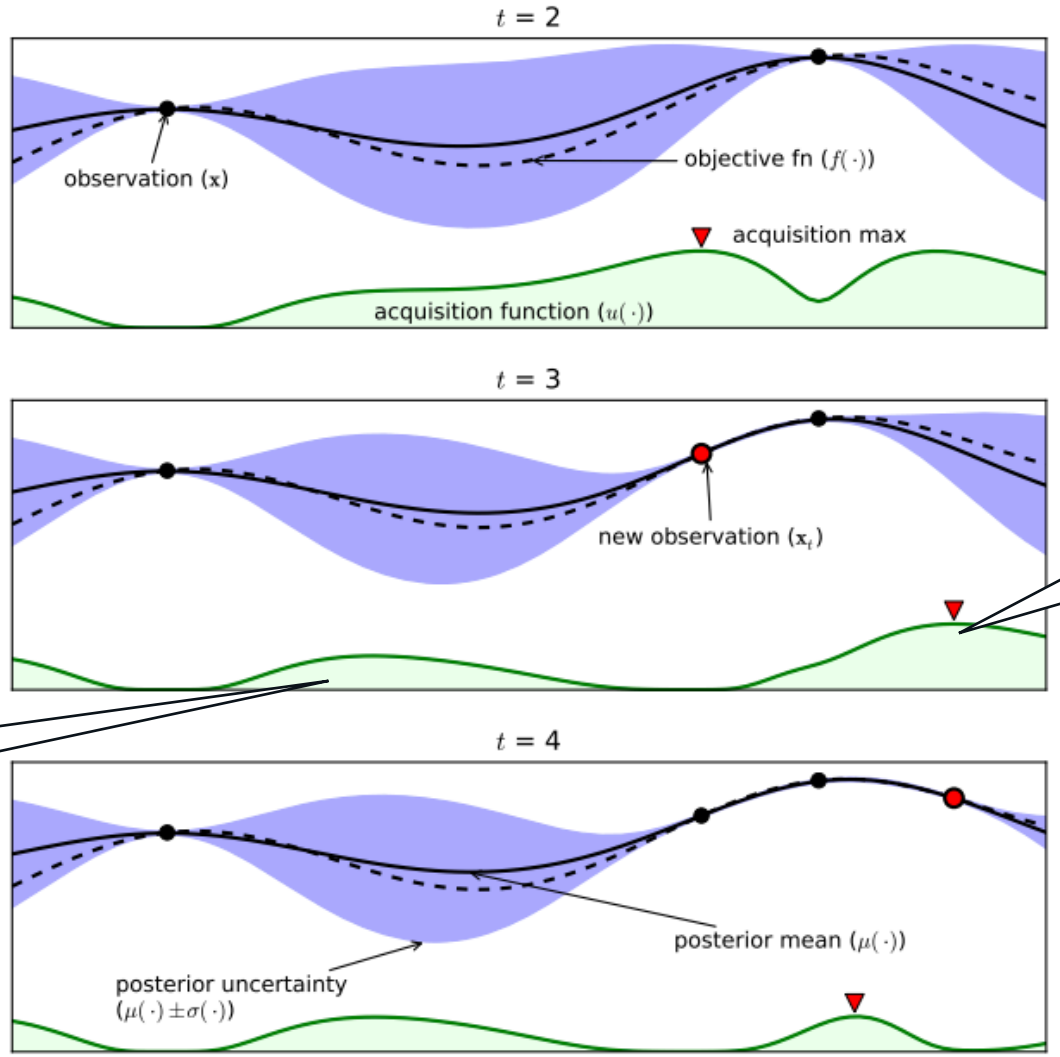
All content copyright © 2019 QuantumBlack, a McKinsey company

..acquisition functions are used to formalize what constitutes a ”best guess”..



All content copyright © 2019 QuantumBlack, a McKinsey company

The expected improvement acquisition trades off *exploitation* of known optimal areas, versus *exploration* of unexplored areas of the loss surface



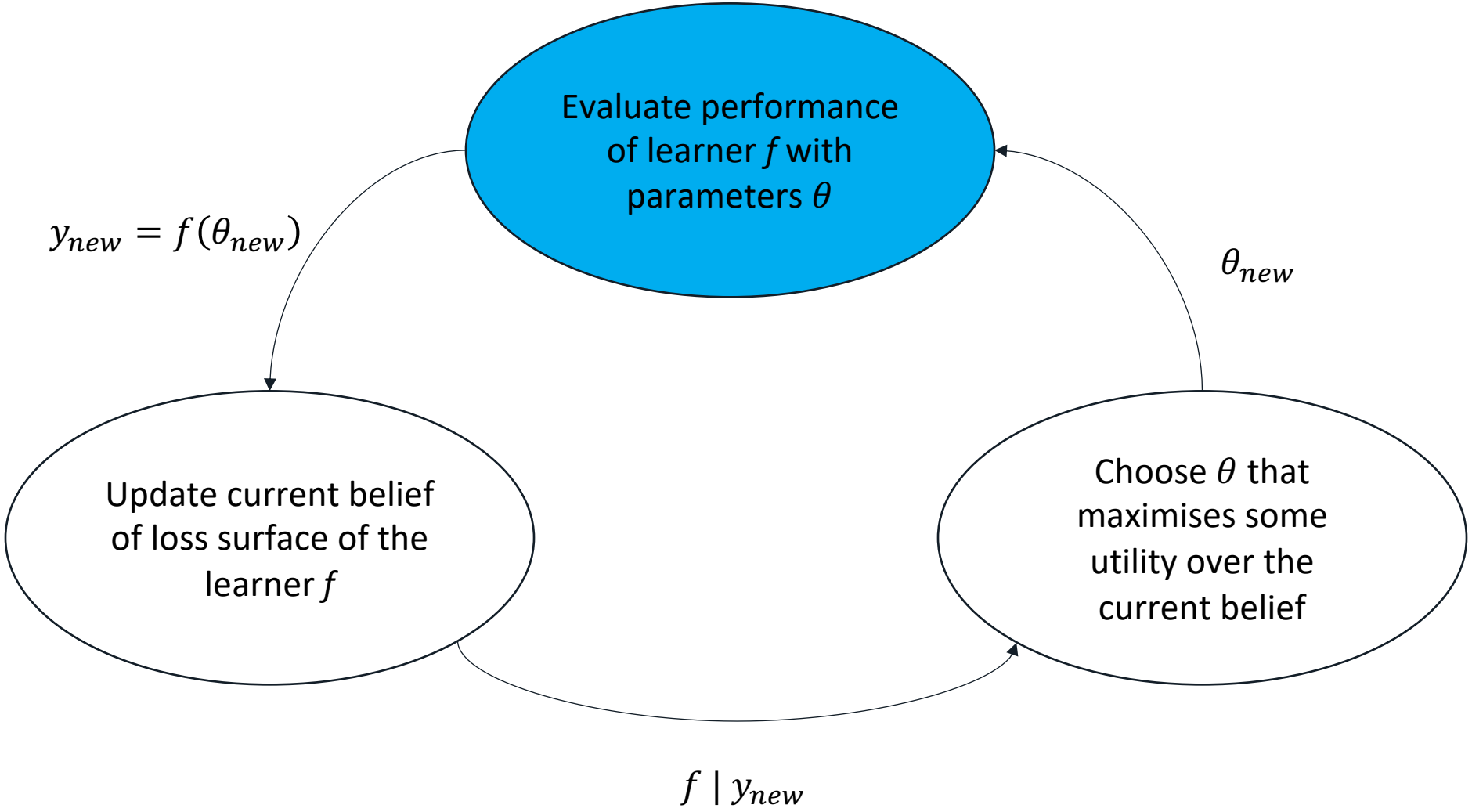
Exploitation of known optimal space

Exploration of unknown space

A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning, *Eric Brochu, Vlad M. Cora, Nando de Freitas*, 2010. https://arxiv.org/abs/1012.2599

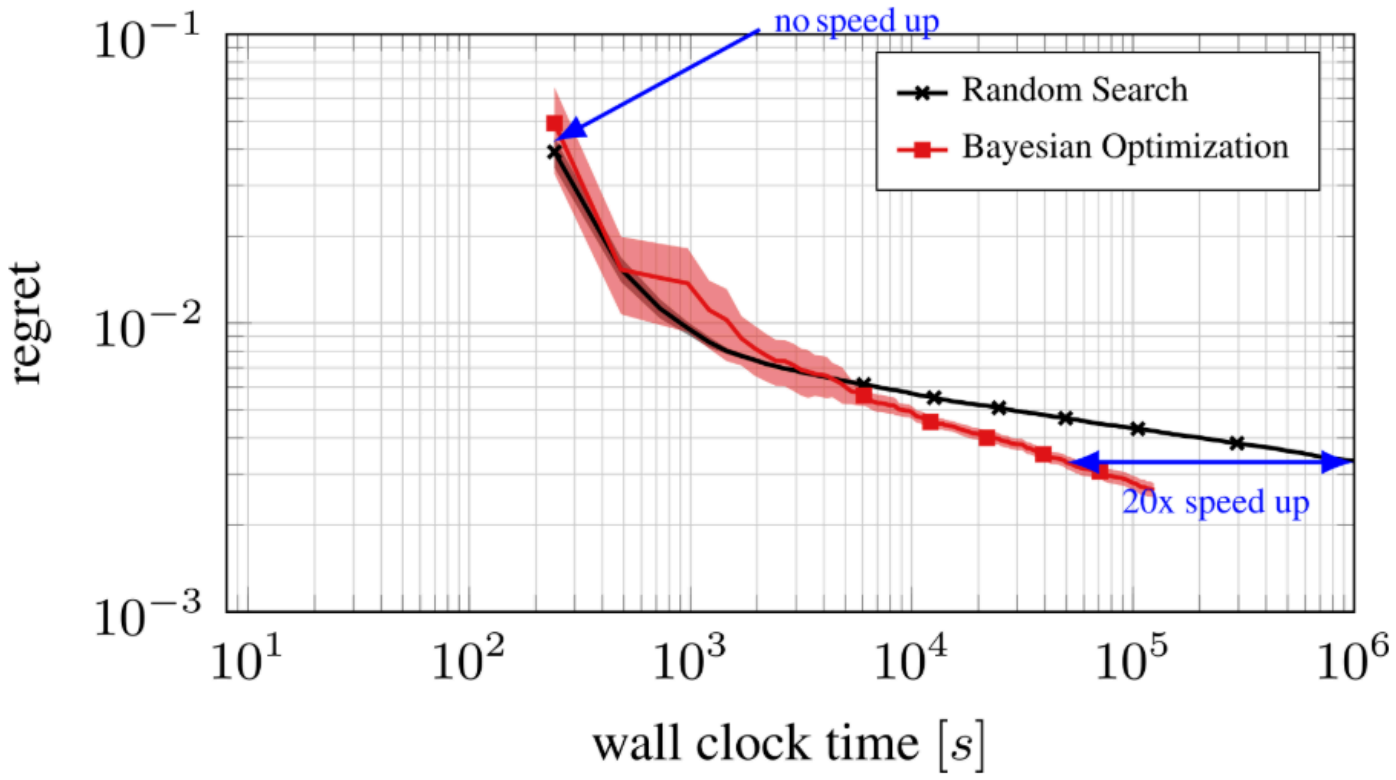
All content copyright © 2019 QuantumBlack, a McKinsey company

.. and finally the performance is computed for suggested parameters



All content copyright © 2019 QuantumBlack, a McKinsey company

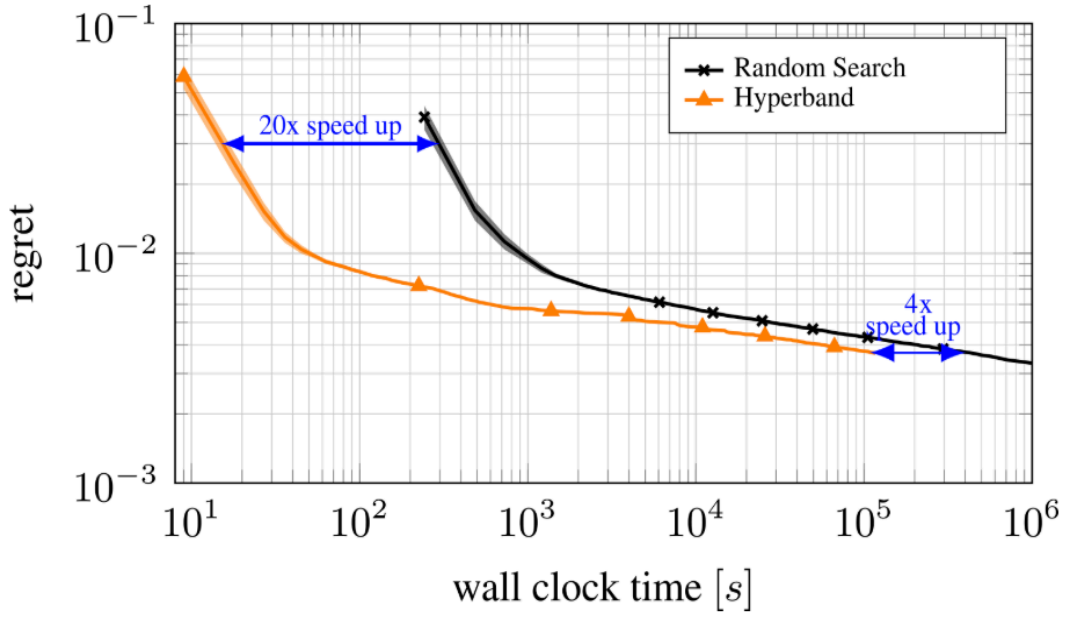
Bayesian optimization is competitive against random search for longer training budgets

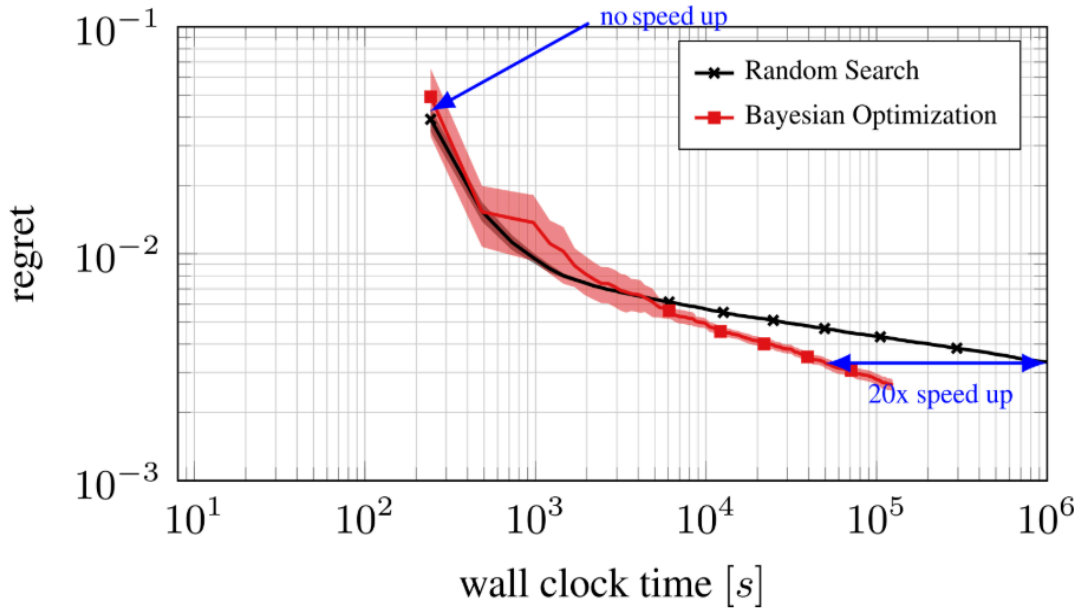


BOHB: Robust and Efficient Hyperparameter Optimization at Scale, ***Stefan Falkner, Aaron Klein, Frank Hutter*****, 2018.** https://arxiv.org/abs/1807.01774

All content copyright © 2019 QuantumBlack, a McKinsey company

Bayesian optimisation works well for larger budgets, whereas hyperband works better for small to medium budgets

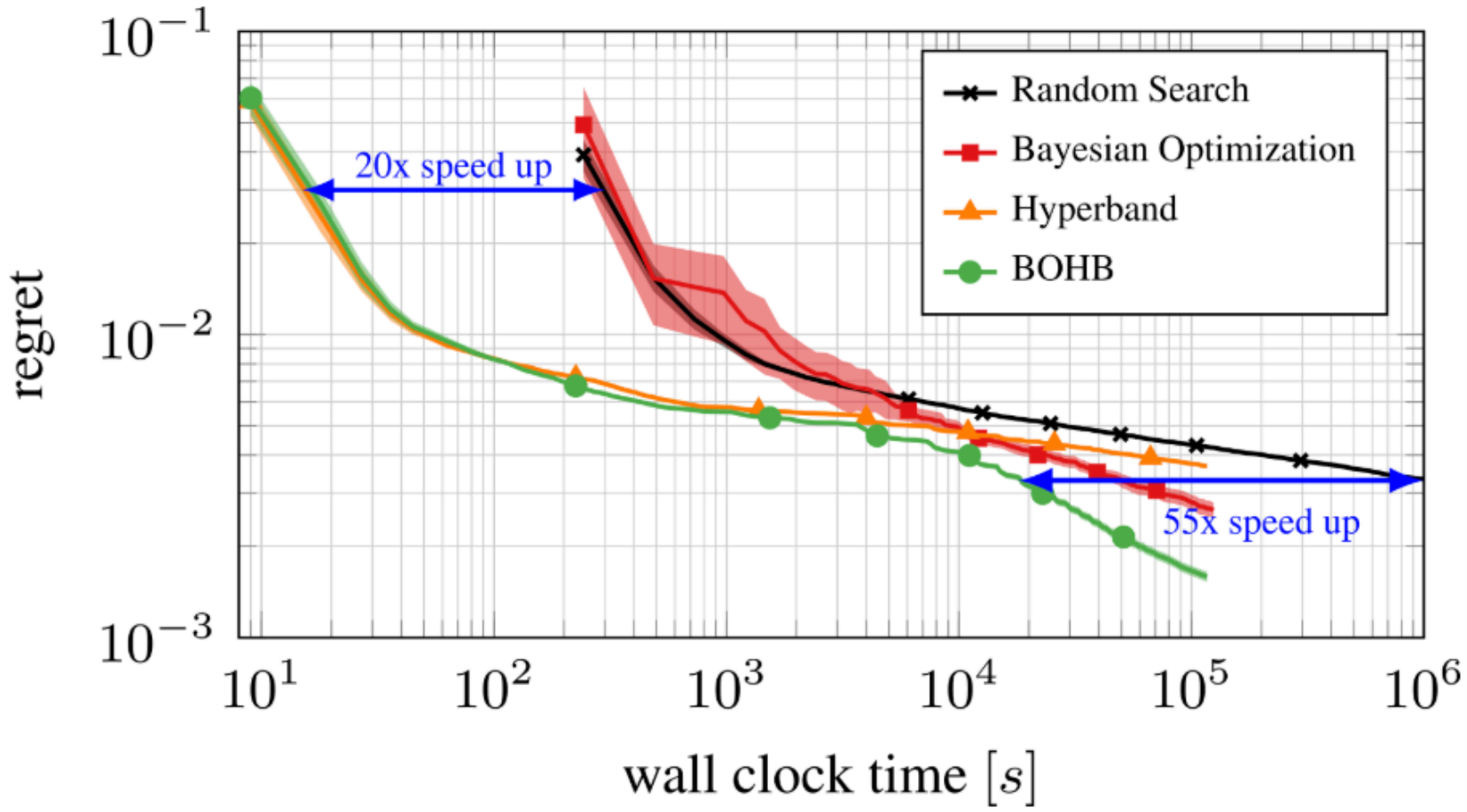




BOHB: Robust and Efficient Hyperparameter Optimization at Scale, *Stefan Falkner, Aaron Klein, Frank Hutter*, 2018. https://arxiv.org/abs/1807.01774

All content copyright © 2019 QuantumBlack, a McKinsey company

BO-HB replaces random search in Hyperband with Bayesian optimisation



BOHB: Robust and Efficient Hyperparameter Optimization at Scale, ***Stefan Falkner, Aaron Klein, Frank Hutter*****, 2018.** https://arxiv.org/abs/1807.01774

All content copyright © 2019 QuantumBlack, a McKinsey company

What HPO method should I use in practice?

- If it is possible to define substantially cheaper version of the performance function $f_{\mathcal{M}}$: **Bayesian optimisation-hyperband (BO-HB)**

- If this is not possible and

  - if all hyperparameters are real-valued and one can only afford a few dozen function evaluations: **Gaussian-process based Bayesian optimisation** (SMAC, Spearmint)

  - if the configuration space is large and/or conditional: **Random forest-based Bayesian optimisation** (hyperopt)

All content copyright © 2019 QuantumBlack, a McKinsey company

Questions?