Week 5 Extension: Automated Hyperparameter Optimization

Tao Huang

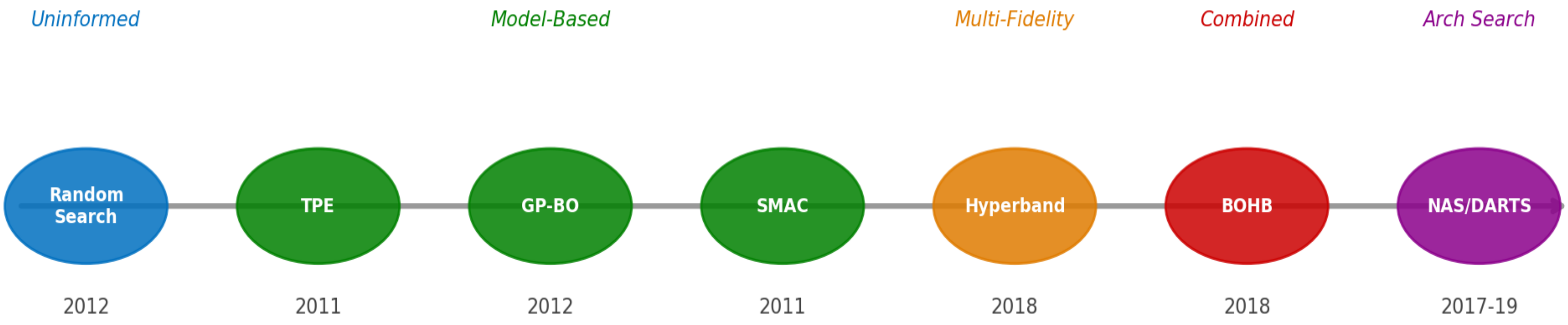
John Hopcroft Center, School of Computer Science, Shanghai Jiao Tong University

https://taohuang.info/cs3317

https://oc.sjtu.edu.cn/courses/89538

From Grid Search to AutoML

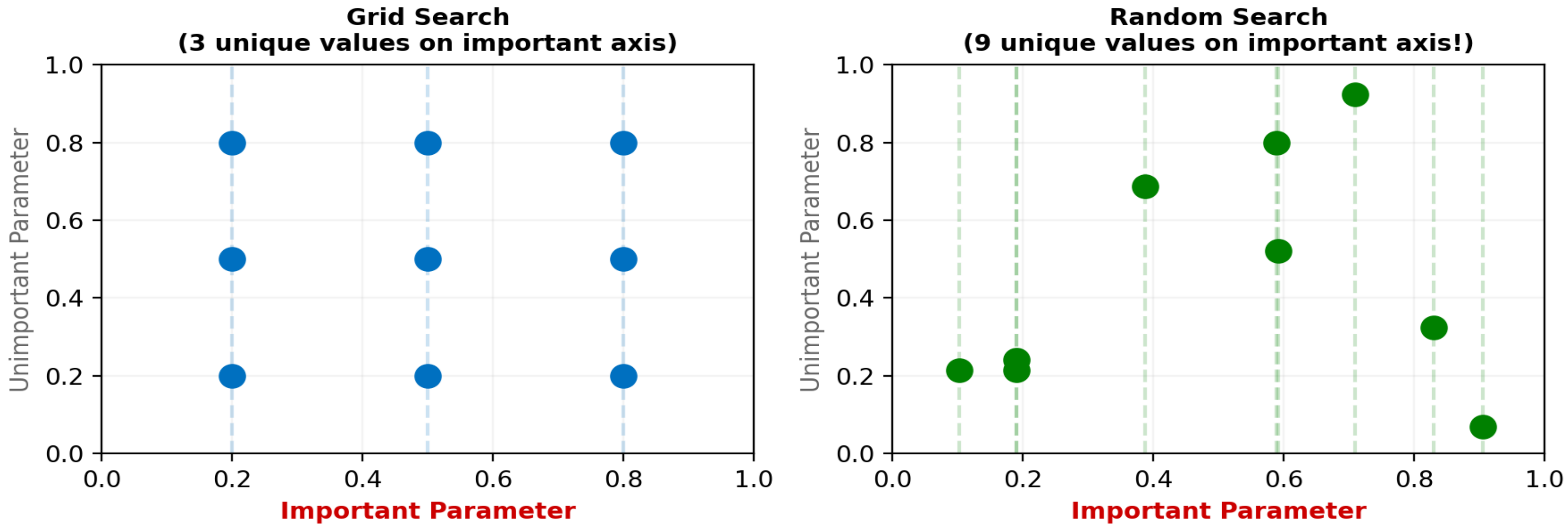
- **Can we do better than Grid Search and Random Search?**
- Key idea: use previous trials to guide the search
- Evolution: uninformed → model-based → budget-aware → architecture search



Revisit: Why Random Search Works

Bergstra & Bengio. Random Search for Hyper-Parameter Optimization. JMLR 2012

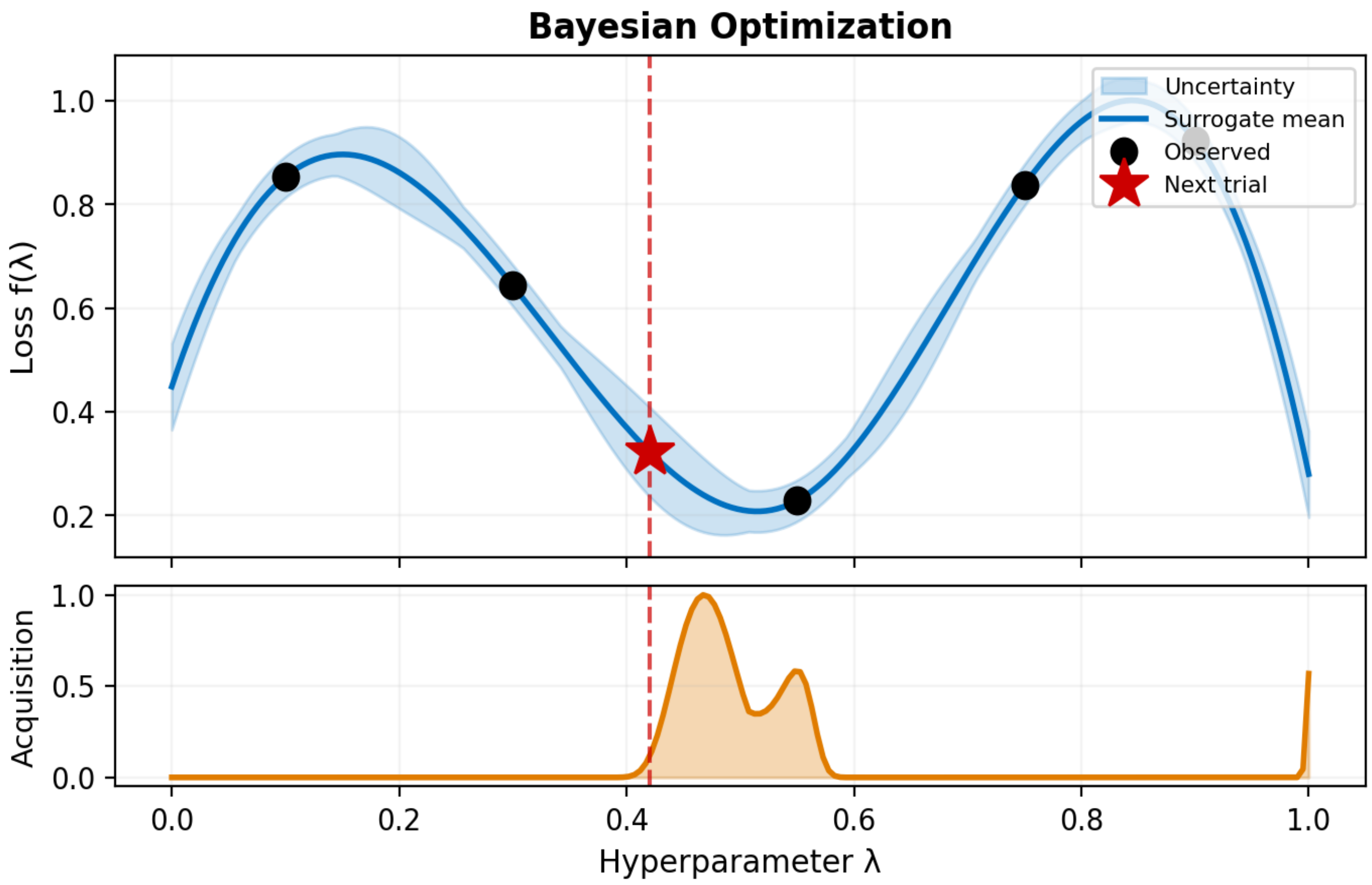
- **Low effective dimensionality: only a few hyperparameters matter**
  - Grid: N^(1/d) unique values per dim
  - Random: N unique values per dim!
- **Same budget, much better coverage**



Bayesian Optimization: The Framework

- **Minimize f(λ) = validation loss (expensive, black-box)**
  - Surrogate: approximates f from past evaluations
  - Acquisition: decides which λ to try next

- Loop: fit surrogate → maximize acquisition → evaluate → repeat



GP-Based Bayesian Optimization

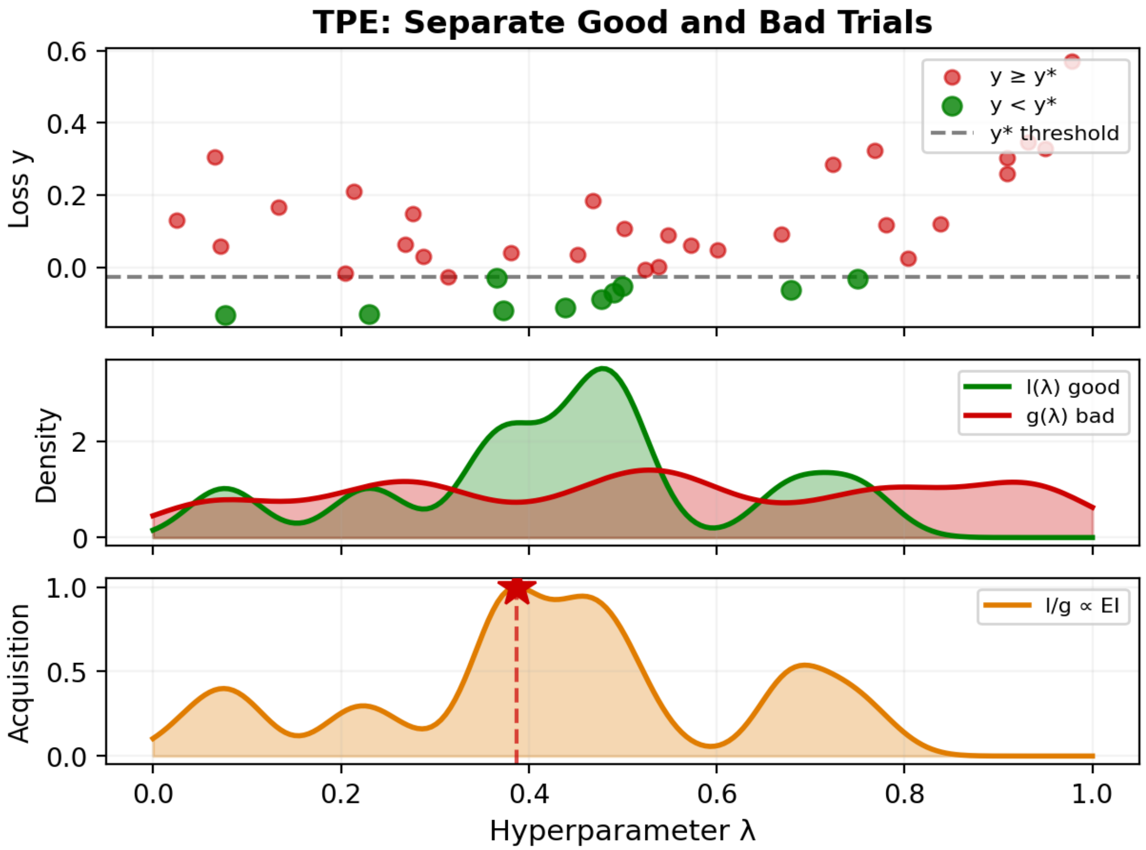
Snoek et al. Practical Bayesian Optimization of ML Algorithms. NeurIPS 2012

- **Surrogate: Gaussian Process (GP)**
  - Predicts mean $\mu(\lambda)$ and variance $\sigma^2(\lambda)$ for untried configs
  - Uncertainty high where few observations exist
- **Acquisition: Expected Improvement (EI)**
  - $EI(\lambda) = E[\max(f\_best - f(\lambda), 0)]$
  - High where loss is low OR uncertainty is high
- Outperformed human experts on CNN and SVM tuning
- Limitation: $O(n^3)$ — struggles above ~1000 observations
- *Library: Spearmint*

Tree-Structured Parzen Estimator (TPE)

Bergstra et al. Algorithms for Hyper-Parameter Optimization. NeurIPS 2011

- **Model two densities:**
  - $l(\lambda) = p(\lambda \mid y < y^*)$ — good configs
  - $g(\lambda) = p(\lambda \mid y \geq y^*)$ — bad configs
- **Maximize $l(\lambda)/g(\lambda) \propto$ EI**
- Advantages: handles categorical,
- scales well, fast computation
- **Default in Hyperopt & Optuna**



SMAC & Surrogate Comparison

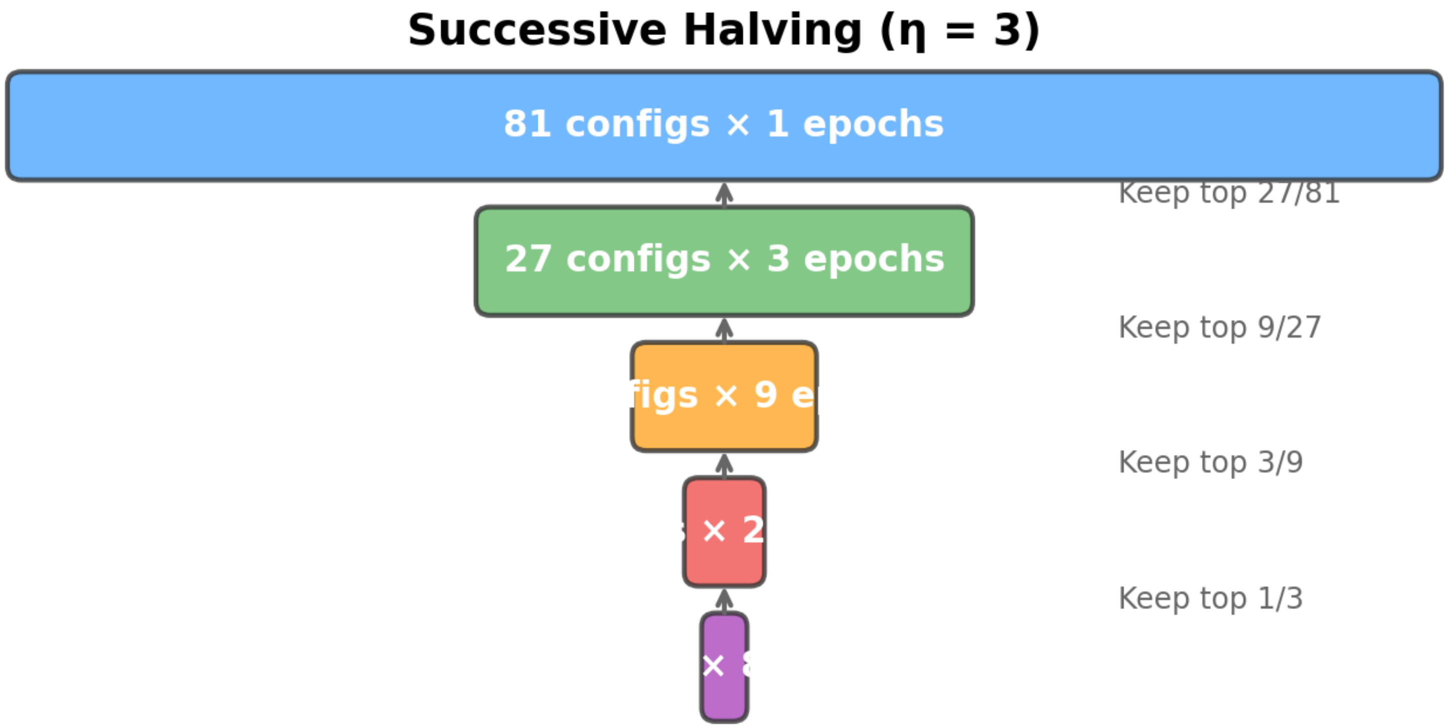
Hutter et al. Sequential Model-based Algorithm Configuration. LION 2011

- Uncertainty = variance across trees. Handles categorical, O(n log n)

Method	Surrogate	Strengths	Scales to
GP-BO (Spearmint)	Gaussian Process	Principled uncertainty	Low-dim continuous
TPE (Optuna)	Density Estimation	Conditional spaces, fast	Mixed types, medium-dim
SMAC (SMAC3)	Random Forest	High-dim, categorical	Mixed types, high-dim

Multi-Fidelity: Stop Bad Runs Early

- **Why train 100 epochs if config looks bad after 10?**
- Successive Halving: start many configs with small budget,
- keep top 1/η, multiply budget by η, repeat



Hyperband

Li et al. Hyperband: A Novel Bandit-Based Approach to HPO. JMLR 2018

- **Run SHA with different aggressiveness levels (brackets)**
- No surrogate — random sampling + early stopping
- Provably ≥ random search, often much faster

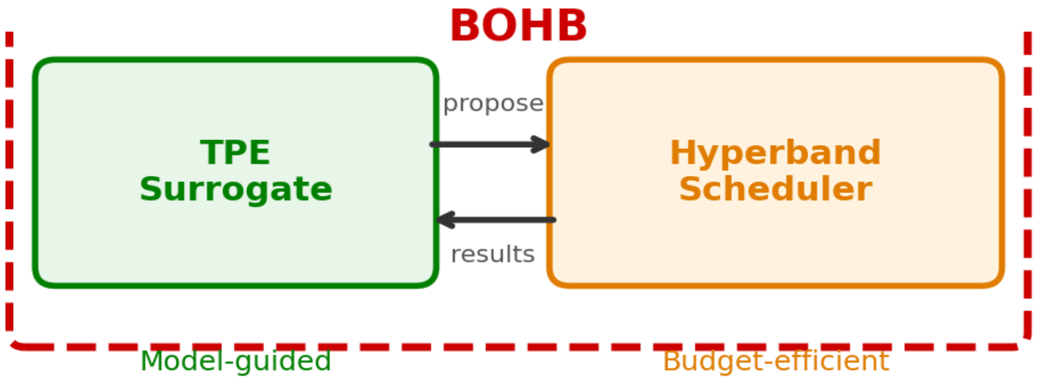
Hyperband Brackets (R=81, η=3)

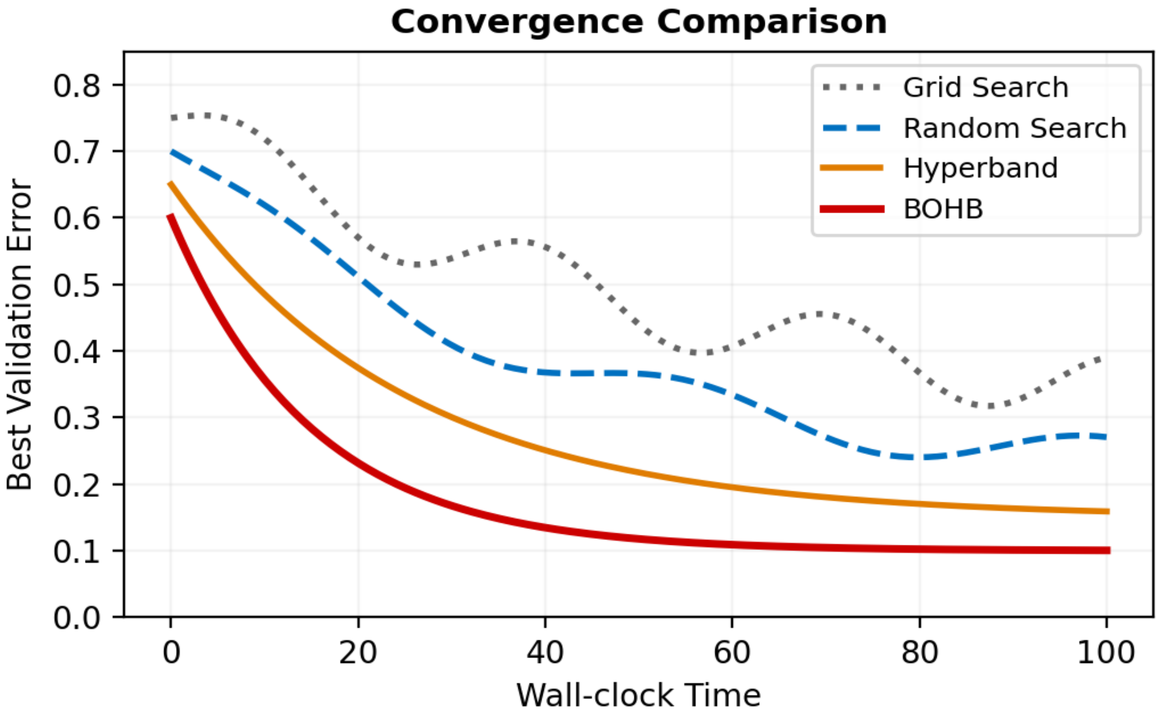
Bracket	n (configs)	min budget	max budget	SHA Rounds
s = 4	81	1	81	5 ← aggressive
s = 3	27	3	81	4
s = 2	9	9	81	3
s = 1	6	27	81	2
s = 0	5	81	81	1 ← conservative

BOHB = BO + Hyperband

Falkner et al. BOHB: Robust and Efficient HPO at Scale. ICML 2018

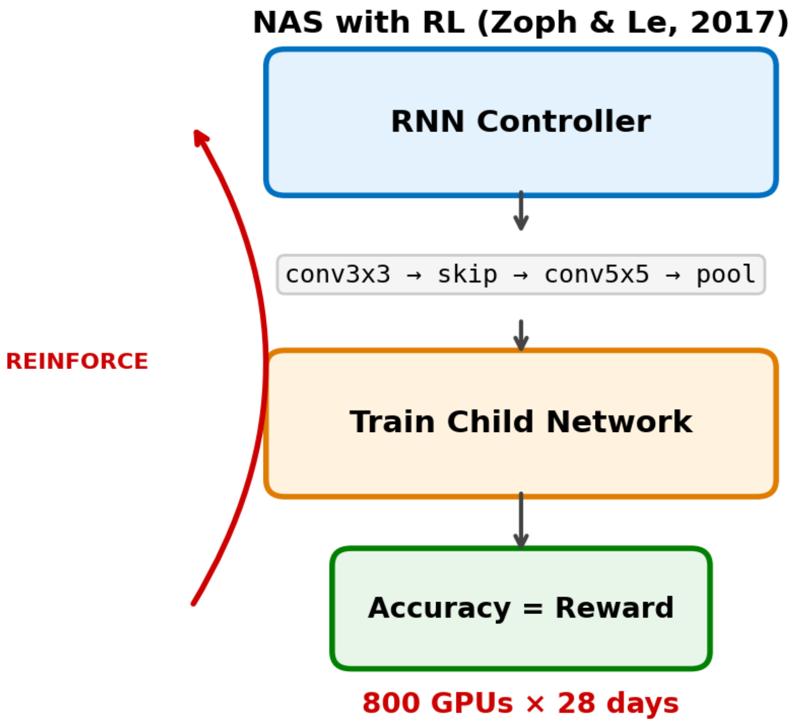
- **Replace random sampling in Hyperband with TPE**
- Best of both: model-guided + budget-efficient

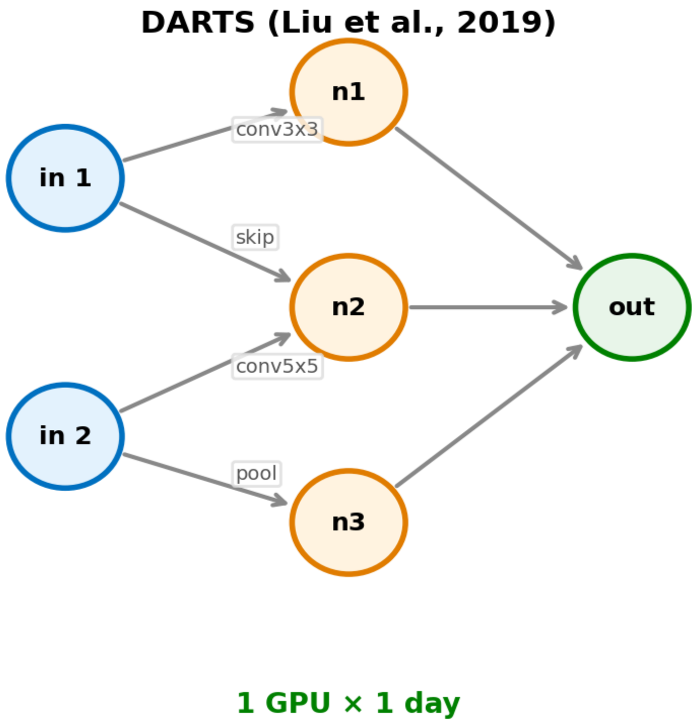




Beyond Hyperparameters: Architecture Search

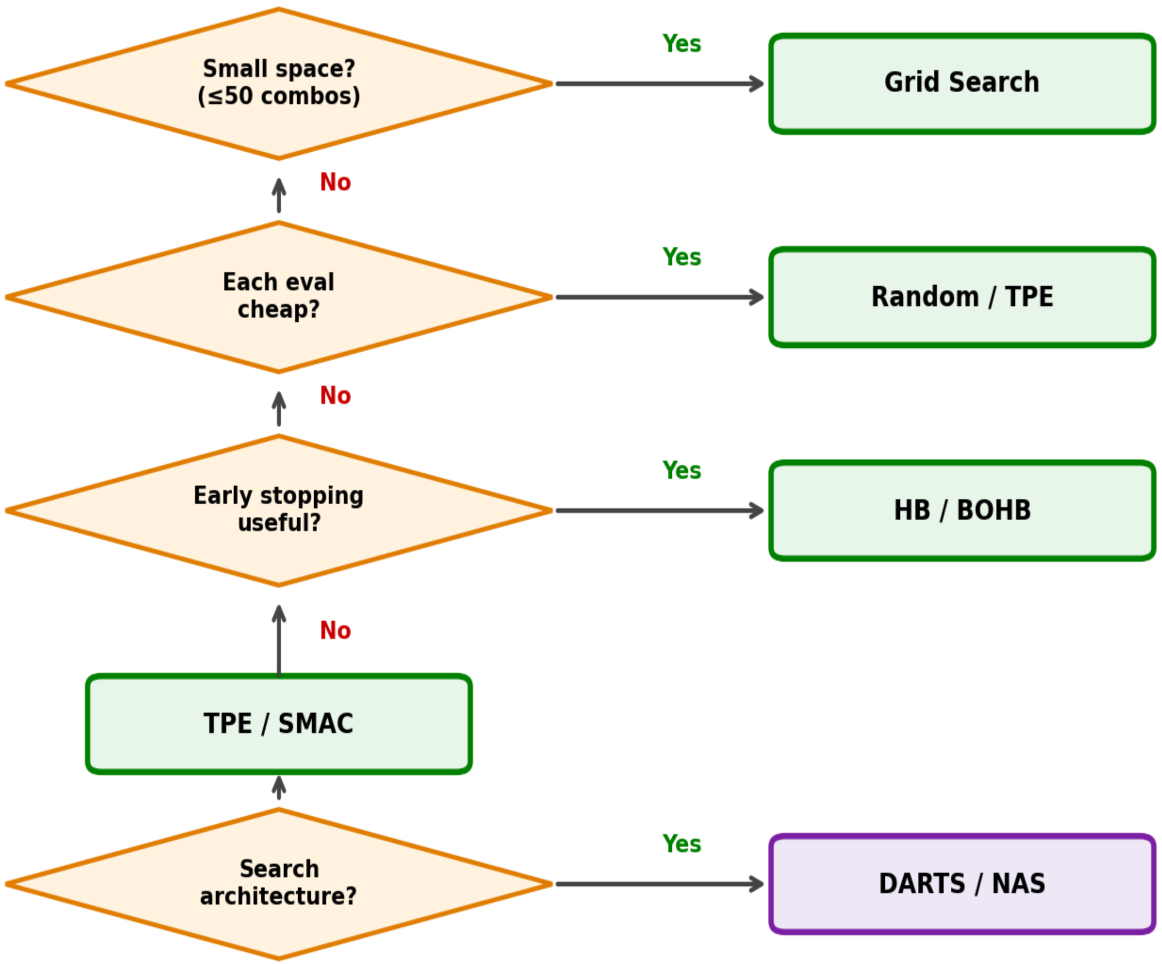
- **Search the model architecture, not just hyperparameters**
- NAS-RL (2017): 800 GPUs × 28 days vs DARTS (2019): 1 GPU × 1 day





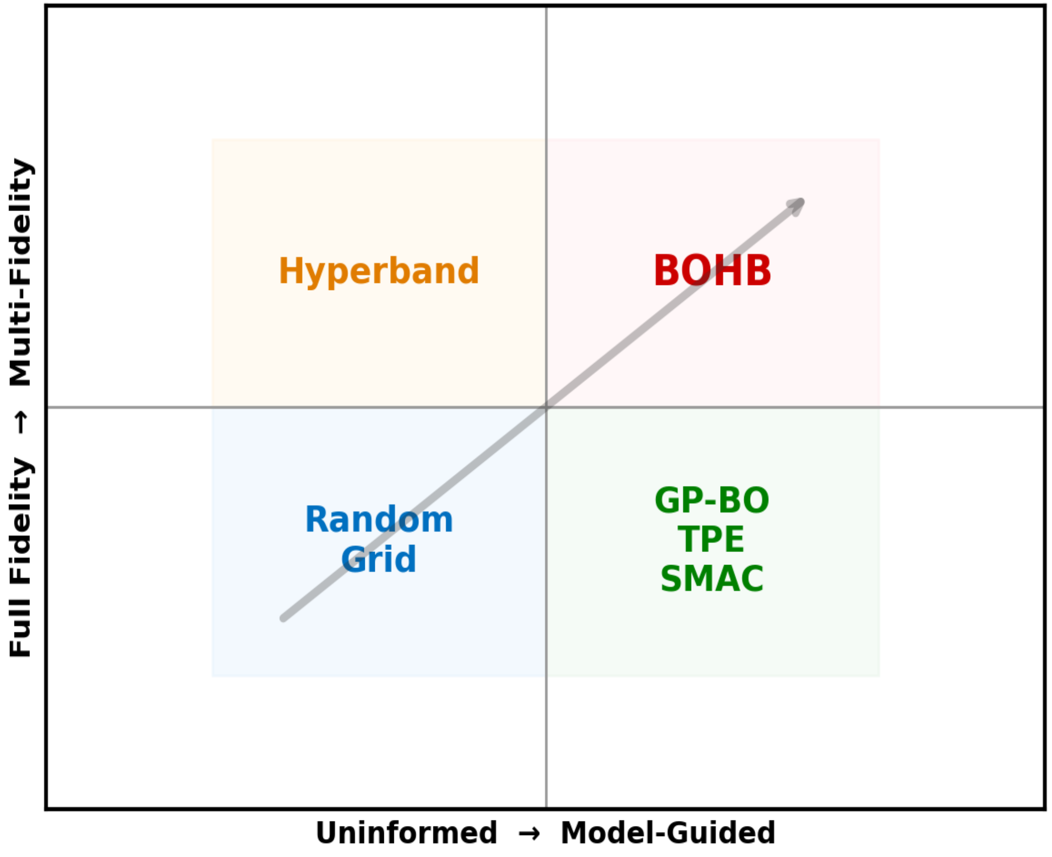
Practical Guide: Which Method?

- **Depends on budget, space size, eval cost**
- Libraries: Optuna (recommended), Ray Tune, SMAC3



Summary: The HPO Landscape

- **Two axes of progress:**
  - Smarter: Random → GP → TPE → SMAC
  - Cheaper: Full training → SHA → Hyperband
- **Combined: BOHB = TPE + Hyperband**
- Frontier: architecture search (NAS, DARTS)
- **Default recommendation: Optuna**



References

Key Papers:

- [1] Bergstra & Bengio. Random Search for Hyper-Parameter Optimization. JMLR 2012
- [2] Bergstra et al. Algorithms for Hyper-Parameter Optimization. NeurIPS 2011
- [3] Snoek et al. Practical Bayesian Optimization of ML Algorithms. NeurIPS 2012
- [4] Hutter et al. Sequential Model-based Algorithm Configuration. LION 2011
- [5] Li et al. Hyperband: A Novel Bandit-Based Approach to HPO. JMLR 2018
- [6] Falkner et al. BOHB: Robust and Efficient HPO at Scale. ICML 2018
- [7] Zoph & Le. Neural Architecture Search with RL. ICLR 2017
- [8] Liu et al. DARTS: Differentiable Architecture Search. ICLR 2019

Further Reading:

- *Feurer & Hutter. Hyperparameter Optimization. AutoML Book, 2019*