# InfoBAX: What inspired this?

Anabel Yong

31st October 2025

**NOTE: There are some miscellaneous concepts on here, which do not make this writeup fully self-contained.**

## 1 Fundamental Overview

Bayesian Optimal Experimental Design (BOED) states:

*Pick the next experiment/design $x$ to maximize expected information gain (EIG) about a target unknown (parameters, a function property, etc.*

This goes back to Lindley's decision theoretic view of experiments[1](maximize mutual information), and the classic survey by Chaloner & Verdinelli[2] that set EIG as a core criterion. Mathematically, if the target is some random quantity $\mathcal{Q}$ and observing $y_x$ is the outcome of running the experiment at design $x$, EIG is:

$$\text{EIG}_x = H[\mathcal{Q}|D_t] - \mathbb{E}_{y_x|D_t}[H|\mathcal{Q}|D_t \cup \{(x, y_x)\}]$$

which is the expected reduction in entropy about $\mathcal{Q}$ after measuring at $x$. InfoBAX [3] will take $\mathcal{Q}$ to be the output of an algorithm run on the unknown function - this is the key jump. In InfoBAX, the authors write the same idea for a chosen algorithm $\mathcal{A}$:

$$\text{EIG}_x = H[\mathcal{O}_A|D_t] - \mathbb{E}_{y_x|D_t}[H|\mathcal{O}_A|D_t \cup \{(x, y_x)\}]$$

### 1.1 Immediate Ancestors: How did InfoBAX come about?

Before InfoBAX, previous Bayesian Optimization papers demonstrated how to **compute/approximate EIG cheaply** by changing *what you learn about.*

1. Entropy Search (ES)[4] targeted information about the $\arg\max x^*$ of the black-box function. It derived mutual information (MI) with respect to $x^*$, sampling from the "posterior over optimizers" and using that to guide evaluations. The mutual information (MI) between two quantities is a measure of the extent to which knowledge of one quantity reduces uncertainty about the other.

2. Predictive Entropy Search (PES)[5] reparameterized the same objective into an equivalent, easier form using the predictive distribution; later PESMO did this for Pareto sets[6].

3. Max-value Entropy Search (MES)[7] simplified further by maximizing information about the maximum value $f^*$ instead of the maximizer, yielding strong performance and much cheaper estimation.

A second strand used mutual information in other Gaussian Processes tasks, Informational Approach to Global Optimization (IAGO)[8] minimized entropy of the minimizer - an early information-thereotic Bayesian optimization method closely related to Hennig and Schuler [4] above. Interestingly, Krause, Singh and Guestrin [9] maximized mutual information to place sensors. This concept here normalized the idea of optimizing information about a property of $f$(predictive field) rather than a single point. Stepwise Uncertainty Reduction (SUR) [10] targeted sets/level sets (e.g. excursion sets), reinforcing the goal-oriented perspective. Build acquisitions around uncertainty in a property of $f$, not just immediate improvement. Finally, Myopic Posterior Sampling (MPS) [11] unified many "goal-oriented" adaptive Design of Experiment (DOE) problems by letting us define a task-specific reward and sampling from the posterior to act myopically, which goes beyond **"finding the maximizer"**.
Additionally, our target is only computable by running a simulator/algorithm, not by a neat tractable likelihood. Approximate Bayesian Computation (ABC), which was first practically demonstrated by geneticists [12], where only simulations providing statistics within some $\epsilon$ of observed values were accepted. The corresponding parameter values approximated samples from the posterior distribution. ABC made it standard by simulating forward and accepting samples close to observations - no closed-form likelihood needed. InfoBAX uses this to condition on algorithm outputs that are defined only via simulating an algorithm on samples of $f$.

# 2 Elegance of InfoBAX

InfoBAX reframes BOED for algorithm outputs: pick $x$ to maximally reduce uncertainty about $\mathcal{O}_A(f)$, the output produced by running some base algorithm $\mathcal{A}$ on the unknown function $f$ (e.g. the top-k set in a finite library, the Pareto front, the root, etc.). The clever part is how to estimate EIG tractably in Equation 1 above. How I have framed this, is that there are two nice big ideas here:

## 2.1 Execution paths as latent variables

When algorithm $\mathcal{A}$ runs on $f$, it generates an execution path $\epsilon_A(f) = \{(z_s, f_{z_s})\}_{s=1}^S$ showing which points it queried and what values it saw. InfoBAX observes that if you knew a plausible execution path, then under a GP, you can compute the posterior predictive at any candidate $x$ conditioned on that path using the

standard "noisy+noiseless" GP conditoning. The intuition here is that the paths act like noiseless pseudo-observations of $f$.
Concretely, with data $D_t$ (noisy) and sampled path $\epsilon_A$(noiseless), the predictive $p(y_x|D_t, \epsilon_A)$ stays Gaussian, with mean/variance obtained by augmenting the Gram matrix with zero-noise blocks for the noiseless points - precisely what the paper writes as the closed-form [3].

### 2.2 Two practical EIG estimators.

ABC view over execution paths: Draw psoterior function samples $f \sim p(f|D_t)$, run algorithm $\mathcal{A}$ on each to get many output/path pairs $(\mathcal{O}_A, \epsilon_A)$, then use an ABC neighbourhood around each simulated output to form approximate samples from $p(\epsilon_A|\mathcal{O}_A, D_t)$. With these, you Monte-Carlo the entropy term in EIG. This is the Stage 1 cache paths, Stage 2 reuse them to score any $x^*$ methodology.
Subsequence (v-variable) estimator. Often the property $\mathcal{O}_A$ determines a small set of function values along part of the execution path (the maximizer's location/value; a level set; top-k values). Then you can push EIG through that lower-dimensional sufficient slice $v$ and compute EIG, in Equation 1, which closed-form under Gaussian Processes, and a great apprxoimation to the full EIG. See Neiswanger, Wang and Ermon [3] for the full discussion.

## Top-K example for any practical experimentation

Suppose $\mathcal{A}$ returns the **top-$k$ items** from a finite set $\mathcal{X}$ (InfoBAX's running example). Define $O_{\mathcal{A}}(f) = K^\star \subset \mathcal{X}$. InfoBAX:

1. Maintains a GP posterior over $f$ from data $D_t$.
2. **Stage 1:** Sample $f^{(j)} \sim p(f \mid D_t)$, run $\mathcal{A}$ on each, recording $(O_{\mathcal{A}}^{(j)}, e_{\mathcal{A}}^{(j)})$. (These are the magenta "simulated outputs," the cached "red dots." in Neiswanger's talk.)
3. **Stage 2:** For any candidate $x$, approximate $\mathrm{EIG}_t(x)$ either
   - **via ABC over paths:** use nearby $O_{\mathcal{A}}$ to approximate samples from $p(e_{\mathcal{A}} \mid O_{\mathcal{A}}, D_t)$ and compute the Monte Carlo entropy reduction; or
   - **via the subsequence estimator** $v$ (e.g., the function values at the items believed to be in the top-$k$), which is closed-form under the GP.

## References

1. Lindley DV. On a Measure of the Information Provided by an Experiment. The Annals of Mathematical Statistics 1956 Dec; 27:986–1005. DOI: 10.1214/aoms/1177728069. Available from: https://doi.org/10.1214/aoms/1177728069

2. Chaloner K and Verdinelli I. Bayesian Experimental Design: A Review. Statistical Science 1995 Aug; 10:273–304. DOI: 10.1214/ss/1177009939. Available from: https://doi.org/10.1214/ss/1177009939

3. Neiswanger W, Wang KA and Ermon S. Bayesian Algorithm Execution: Estimating Computable Properties of Black-box Functions Using Mutual Information. 2021. arXiv: 2104.09460 [stat.ML]. Available from: https://arxiv.org/abs/2104.09460

4. Hennig P and Schuler CJ. Entropy Search for Information-Efficient Global Optimization. 2011. arXiv: 1112.1217 [stat.ML]. Available from: https://arxiv.org/abs/1112.1217

5. Hernández-Lobato JM, Gelbart MA, Hoffman MW, Adams RP and Ghahramani Z. Predictive Entropy Search for Bayesian Optimization with Unknown Constraints. 2015. arXiv: 1502.05312 [stat.ML]. Available from: https://arxiv.org/abs/1502.05312

6. Hernandez-Lobato D, Hernandez-Lobato J, Shah A and Adams R. Predictive Entropy Search for Multi-objective Bayesian Optimization. *Proceedings of The 33rd International Conference on Machine Learning*. Ed. by Balcan MF and Weinberger KQ. Vol. 48. Proceedings of Machine Learning Research. New York, New York, USA: PMLR, 2016 :1492–501. Available from: https://proceedings.mlr.press/v48/hernandez-lobatoa16.html

7. Wang Z and Jegelka S. Max-value Entropy Search for Efficient Bayesian Optimization. 2018. arXiv: 1703.01968 [stat.ML]. Available from: https://arxiv.org/abs/1703.01968

8. Villemonteix J, Vazquez E and Walter E. An Informational Approach to the Global Optimization of Expensive-to-Evaluate Functions. Journal of Global Optimization 2009; 44:509–34. DOI: 10.1007/s10898-008-9354-2. Available from: https://doi.org/10.1007/s10898-008-9354-2

9. Krause A, Singh A and Guestrin C. Near-Optimal Sensor Placements in Gaussian Processes: Theory, Efficient Algorithms and Empirical Studies. Journal of Machine Learning Research 2008; 9:235–84. Available from: http://jmlr.org/papers/v9/krause08a.html

10. Chevalier C, Bect J, Ginsbourger D, Vazquez E, Picheny V and Richet Y. Fast Parallel Kriging-Based Stepwise Uncertainty Reduction with Application to the Identification of an Excursion Set. Technometrics 2014; 56:455–65. DOI: 10.1080/00401706.2013.860918. Available from: https://doi.org/10.1080/00401706.2013.860918

11. Kandasamy K, Neiswanger W, Zhang R, Krishnamurthy A, Schneider J and Poczos B. Myopic Posterior Sampling for Adaptive Goal Oriented Design of Experiments. *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Chaudhuri K and Salakhutdinov R. Vol. 97. Proceedings of Machine Learning Research. PMLR, 2019 Sep :3222–32. Available from: https://proceedings.mlr.press/v97/kandasamy19a.html

12. Pritchard JK, Seielstad MT, Perez-Lezaun A and Feldman MW. Population Growth of Human Y Chromosomes: A Study of Y Chromosome Microsatellites. Molecular Biology and Evolution 1999; 16:1791–8