Published in final edited form as:
Curr Opin Behav Sci. 2019 October ; 29: 117–126. doi:10.1016/j.cobeha.2019.07.008.

What is optimal in optimal inference?

Gaia Tavoni[a,b], **Vijay Balasubramanian**[a,b], **Joshua I. Gold**[a]

[a]Department of Neuroscience, University of Pennsylvania, Philadelphia, PA 19104

[b]Department of Physics and Astronomy, University of Pennsylvania, Philadelphia, PA 19104

Abstract

Inferring hidden structure from noisy observations is a problem addressed by Bayesian statistical learning, which aims to identify optimal models of the process that generated the observations given assumptions that constrain the space of potential solutions. Animals and machines face similar “model-selection” problems to infer latent properties and predict future states of the world. Here we review recent attempts to explain how intelligent agents address these challenges and how their solutions relate to Bayesian principles. We focus on how constraints on available information and resources affect inference and propose a general framework that uses benefit(accuracy) and accuracy(cost) curves to assess optimality under these constraints.

Introduction

We continuously gather noisy and incomplete information to make inferences about past, present, and future states of the world. Classical Bayesian theory provides a probabilistic framework to identify optimal solutions to this problem [1]. However, Bayesian approaches to inference typically rely on immediate access to all available and appropriate sources of information and unlimited time and resources to carry out the necessary computations. These prerequisites can be met only partially for inference problems faced in the real world, leading to questions about the applicability of Bayesian principles for understanding how the brain and other intelligent agents solve these problems.

Here we discuss recent findings that are beginning to provide a more nuanced understanding of how optimality can be defined in conditions that are subject to real-world constraints, which include limits on the available information and the time and other resources needed to carry out the inference process. We focus on identifying the conditions under which inference strategies employed by animals and machines are consistent with Bayesian principles and those in which they deviate from them. We show that the effectiveness of strategies under real-world constraints could be assessed by maximizing benefit/cost, where realistic benefit(cost) curves can be defined by combining two separate components, corresponding to benefit(accuracy) and accuracy(cost) curves. We argue that this approach can, in principle, be used to explain: (1) the large variation in cost and complexity of

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

inference strategies used for different tasks and conditions, and (2) why individuals sometimes choose inference strategies that are seemingly sub-optimal compared to the Bayesian solution.

Constraints on accessible information

Animals face the fundamental problem of inferring the model that best explains noisy observations, given information from *a priori* beliefs and ongoing observations. Both of these kinds of information can be limiting. For example, prior knowledge may or not specify the functional form of the process that generated the observations. Likewise, sampling of the environment may be constrained by how much and what kind of data can be gathered. Bayesian theory predicts how these constraints affect the speed of learning and the complexity the learned model. Below we explain these predictions and compare them with animal behavior in three scenarios: (1) the underlying generative process is known, leaving uncertainty only about its parameters; (2) the underlying process is unknown but lies within a specified hypothesis space of processes; (3) there is no prior knowledge of the hypothesis space for the generative process.

Uncertainty about model parameters

When the functional form of the model is known, Bayesian methods seek to infer the optimal parameter values by maximizing the posterior distribution of the parameters θ given the observed data x (where θ and x can be vectors). This distribution is obtained through Bayes' rule: $p(\theta/x) \propto p(x/\theta)p(\theta)$, where $p(x/\theta)$ is the likelihood of the data x in the model parameterized by θ, and $p(\theta)$ reflects a priori expectations about the parameter values and further constrains the hypothesis space.

Strong sampling, weak sampling, and the "size principle": In general, as more data are observed, the Bayesian posterior converges to the likelihood (and departs from the prior). If sampling is "strong", i.e., if samples are drawn independently at random from the model to infer, this convergence occurs exponentially fast in the number of samples. Consider a special case in which the different parameters θ index categories such that each θ describes a probability distribution, $p(x/\theta)$, that is nonzero in a subset x_θ of the sample space. The elements of x_θ are then examples of the category, and it is possible that the x_θ overlap or intersect. In this setting, if examples are drawn independently at random from a given underlying category (strong sampling), the likelihood of seeing examples under that category is inversely proportional to the number of items contained in that category. As a result, the Bayesian posterior concentrates exponentially quickly in the number of observed examples on the smallest and most specific category from which the examples could be drawn. This effect, known as the "size principle", has been proposed to explain the striking human ability to learn categories from a few examples [2, 3, 4, 5]. In contrast, if sampling is "weak", the relation between the sample distribution and the true distribution of items in the category is not known, because of which all categories compatible with observed samples will be equally likely. In this case, the posterior probability does not concentrate on the most specific category. Human categorization sharpens with increasing data but at variable speeds and generally not as fast as the size principle would prescribe. This result has been interpreted in terms of different assumptions that people make about how observations are generated, usually compatible with a mix of strong and weak sampling [3, 4, 5].

Hierarchical Bayesian Models: Other sampling effects are evident in more complex problems addressed by Hierarchical Bayesian Models (HBMs) [6, 7]. HBMs solve inference problems in which the prior probability is split on multiple levels: typically, each level defines the probability of certain parameters conditional on other parameters, whose probability is defined at the immediately higher level. The distribution at the top level can be thought of as an "innate prior". Thus, these models represent increasingly abstract knowledge, for example about classes of categories, along a hierarchy, and can learn at multiple levels of abstraction simultaneously [8]. This property has made HBMs successful as models of language learning [3, 8, 9, 10, 11, 12], conceptual learning [7, 13, 14], reinforcement learning [15], and adaptive inference of dynamic internal states [16, 17, 18, 19]. HBMs make quantitative predictions about how the speed of learning structure in the data at different levels of abstraction is determined by both the sampling process and the structure of the data. For example, abstract knowledge can be acquired faster than specific knowledge: (1) when sampling interleaves examples of different categories [6, 13]; (2) if independent, random sampling results in a skewed distribution of categories, because the incidence of low frequency categories supports inference of the existence of novel categories [13]; and (3) if the features defining a category have little variance among the items of the category [20]. These kinds of sampling constraints are also likely to play critical roles in inference problems in which there is hierarchical or complex structure in the temporal sequences of observations, for which both Bayesian and other solutions require sampling over potentially many different timescales that provide relevant information about volatility of the environment [16, 17, 21, 22, 23, 24, 25, 26, 27, 28].

Uncertainty about the model form in a constrained hypothesis space

When the functional form of the model generating the observed data is unknown, but the set of plausible models is constrained, Bayesian theory can again be used to select the appropriate model form. In this case, the posterior for a certain model form is defined by integrating over the manifold of all the model probability distributions that have the same functional form f: $p(f \mid x) \propto p(f) \int d\theta p(x|\theta, f) p(\theta|f)$. Assuming a uniform prior over distributions (called Jeffreys' prior), the log-posterior can be expressed asymptotically as a sum of terms of decreasing importance with increasing number N of data points: the leading-order term ($\mathcal{O}(N)$) is the log-likelihood of the data under the optimal model and represents fitting accuracy, and the lower-order terms ($\mathcal{O}(\log N)$ and $\mathcal{O}(1)$) correspond to a measure of statistical simplicity of the model family, which is important to prevent overfitting [33, 34]. These competing quantities produce tradeoffs between the model complexity and fitting accuracy in Bayesian model selection as the number of sampled observations changes.

Bias-variance tradeoffs as a balance between fitting accuracy and complexity: The asymptotic expansion of the Bayesian posterior formalizes the biasvariance trade-off, such that simple models tend to underfit the data and thus yield biased results with low variance across resamplings from the true model, whereas complex models tend to fit the data but overfit the noise and thus yield lower bias but higher variance (Fig.1A,B,C). The optimal model family, which can maximally generalize to new data sampled from the true generative process, typically has intermediate statistical complexity. The tradeoff between fitting accuracy and complexity is sensitive to amount of data. When the amount of data is small, complexity and accuracy compete on an equal basis in model selection. With increasing amounts of data, model accuracy dominates model selection, driving the inference of more complex models that better explain structure in the data. Recent studies on perception [35] and human inference in simple sensorimotor (curve-fitting) tasks [29, 30] have shown, among models of different complexities and equivalent fitting accuracy, people prefer the simplest option, which makes fewer assumptions about the generative process and generalizes better. However, as more data accumulate individuals tend to infer more complex models, as expected from principles of Bayesian inference [36].

Complex model selection: Even if the set of plausible models or hypothesis space is known, it can be highly complex, with many correlated hypotheses (i.e., partially overlapping model families in the parameter space) and non-uniform prior densities. Under these conditions, consistent Bayesian model selection requires that the different priors match over the shared structure of the models. However, this constraint can lead to an interference effect on model selection given by the particular hypothesis space that is considered: for a fixed generative process, if a model is preferred to another one within a given hypothesis space, this preference is not guaranteed to be preserved when the same models are embedded in a different hypothesis space [37]. It would be interesting to understand if and how these effects relate to interference, measurement-order, and other contextual effects found in psychology [38, 39, 40, 41, 42, 43].

Uncertainty about the hypothesis space

In real-world statistical-inference problems, there may be so much uncertainty about which hypotheses should be considered that identifying a set of options that is likely to contain the true generative model is not feasible [44]. Ideally, the goal of model selection in these cases is to find a model within a given hypothesis space that is as simple as possible and yet comes close to the true data-generating distribution, with closeness quantified in this tradeoff by the log-likelihood of the observed data [33]. In this setting, we are not looking for the “optimal” model but simply the best choice in the set that is feasible to consider.

Ecological rationality – simple though biased models work better: Real-world problems in which the optimal model cannot be found because relevant information about the structure of the environment is lacking are the focus of “ecological rationality,” which aims to determine which models or strategies are better than others and in which environments [45, 46, 47, 48, 49, 50, 51, 52]. In general, these studies have highlighted a “less-is-more” effect, whereby simple fast-and-frugal heuristics or experience-based intuition, which may have evolved to meet recurrent needs in animal evolution, outperform complex statistical models that require more computations, evidence, and assumptions [48, 49, 52, 53, 47]. This effect is consistent with expectation from Bayesian theory [33] and can also be interpreted as a consequence of the bias-variance trade-off. Specifically, when uncertainty is high because of an underconstrained hypothesis space or because of excessive noise or instability of the environment, subjects often opt for simple strategies that are inflexible (high bias) but marginally affected by noise (low variance). These simple strategies can match, or in some cases surpass, the performance of more complex strategies that minimize bias at the expense of higher variance [54, 48]. The simple strategies may then form building blocks for more complex and adaptable strategies that are needed only under particular conditions, such as when the environment is moderately unstable [55, 46, 56].

Constraints on resources

Bayesian inference often involves complex manipulations of probabilities requiring a long time to perform, extensive memory, and/or substantial investment of computational resources [57, 58, 7, 59]. In practice, however, computational resources are limited and inferences must be made quickly to be useful. Thus, recent work has sought to understand Bayesian optimality in the presence of resource and time constraints.

Inference with limited time

Animals: Even when sufficient information for optimal Bayesian inference is available in the environment or from memory, the brain needs time to extract and process this information, implying a speed-accuracy trade-off (Fig.1D,E,F) [60, 61, 62, 63, 64, 65, 66, 67]. For the kinds of two-alternative forced-choice (TAFC) tasks used commonly to study perception, memory, and value-based decision-making, the optimal trade-off that maximizes expected accuracy for a given processing time is implemented by the drift-diffusion model (DDM). The DDM selects the alternative with the highest Bayesian posterior given the portion of the evidence that can be processed or accumulated in the available time [68]. These computations can be thought of as Bayesian inference constrained by limited time for sequential information processing, and are consistent with both behavioural and neural responses in these tasks [62, 63, 64, 66, 69, 70]. The optimal speed-accuracy trade-off implemented by the DDM can also be seen, using a physics perspective, in terms of a variational principle that trades off a negative "energy" (representing performance gains) against an information-theoretic distance between the prior and the posterior probability distributions (representing information-processing costs) over the possible options. This interpretation is part of a new statistical theory of bounded, rational decision-making [71].

Machines: Trade-offs between accuracy and computational speed are also prominent in machine-learning applications, as exemplified by "Anytime" algorithms that are designed to generate increasingly accurate solutions to specific problems the longer they run (Fig.1G,H,I) [58, 72]. Anytime algorithms can be interrupted at any time and still produce a valid solution, which gives them the flexibility to solve inference and other decision-making tasks under variable time constraints. They are also well suited to complex problems in which exact solutions are often hard to find and require substantial computational resources and time. Recent examples include the problem of real-time, efficient distribution of power [73]; large-scale optimal web service composition [74]; ranked information retrieval [75]; approximate resolution in limited time of NP-hard problems, such as the identification of common patterns in strings of data [76] and the problem of heterogeneous multirobot coordination [77, 78, 79]; the SLAM (simultaneous localization and mapping) problem for robot navigation [80]; the related motion-planning problem [81]; and pattern recognition [32] and classification [82]. It would be interesting to understand whether such algorithms have a neural or cognitive realization.

Inference with limited computation and memory

Animals: Even without a constraint on processing time, the optimal Bayesian solution might be unachievable because of limitations of computational resources and memory capacity. This idea has led researchers to seek alternative algorithms that the brain might use to solve statistical inference tasks subject to resource constraints. This problem has been studied in some detail in the context of inference in dynamic environments, where latent states, like the source of an uncertain reward, change in time with unknown and possibly time-varying volatility. When the latent states drift continuously in time, they can be inferred with limited memory and computational costs using an algorithm that is derived from a variational-Bayesian approximation of the posterior distribution of the states in a hierarchical Bayesian model [16, 17]. This algorithm implements Markovian equations with dynamic learning rates that weigh prediction errors based on the amount of uncertainty in the observations [16, 17]. These equations closely resemble classical reinforcement-learning heuristics and might therefore be implementable in the brain [19, 18]. Several approximate Bayesian algorithms have also been proposed to explain how the brain might infer latent states that do not drift but undergo sudden change points. These algorithms include particle filters, which reduce the computational and memory costs by approximating the Bayesian posterior using a small set of Monte Carlo samples that is updated as new observations are made [22, 83, 84, 85, 86, 87, 88]; approximate Bayesian models in which the memory load is reduced by forgetting or exponentially discounting past information [55, 24, 89, 90, 91, 92]; low-dimensional approximations of Bayesian models that can infer dynamic discrete-valued states and could be implemented by neural networks with plausible plasticity rules [25]; integrate-and-fire neuron models [28] and leaky evidence-accumulation models [27] that can infer dynamic binary states. All of these models approximate optimal solutions with different levels of accuracy and different computational costs and, in many tasks and conditions, match human behaviour more closely than exact Bayesian models [89, 90, 91, 83, 84, 85, 87, 88]. Ongoing work is assessing how some of these solutions, and the cognitive operations that they represent, are related to one another, the quantitative form of the cost-accuracy trade-off that emerges from them, and how this trade-off can be optimized in different environments [55].

Machines: Trade-offs between accuracy and computational costs are also studied in machine learning. For example, in the field of deep neural networks (DNNs), there is a growing demand for efficient machine learning in devices with limited memory and computational power [93, 94, 95, 96, 97, 98, 99, 100]. These trade-off have been characterized in terms of different performance metrics, including: (1) “information density,” or the accuracy-per-parameter ratio that takes into account architectural complexity (the parameters used) but not computational complexity (the operations performed) [101]; (2) “NetScore,” which gauges accuracy relative to both architectural and computational complexity [102]; and (3) predictive accuracy per input compression, and its distance to the “information-bottleneck” theoretical limit, which defines the maximum predictive accuracy that can be achieved for any given level of compression of the input [103, 104, 105]. These and other metrics are being used to compare DNNs and elucidate the specific cost-accuracy trade-off that they must navigate. Large scale comparisons of DNNs for object detection [106] and image classification [101, 93] have shown, for instance, that different DNNs lie at different points of a specific monotonic trade-off between accuracy and computational complexity: as the amount of computation increases, accuracy also increases but at progressively smaller rates. This phenomenon also applies to the theoretical information-bottleneck limit, which predicts diminishing returns in performance with increasing complexity of input encoding [104]. Identification of upper bounds on accuracy given different constraints on memory and computational costs and comparison of these bounds with current DNN performance can be useful to guide selection of the DNNs that best match the constraints of each application [101, 102, 107] and to develop better architectures and training algorithms [103].

What is optimal in optimal inference?

Classically, optimal Bayesian inference balances prior knowledge and ongoing observations to identify the model with maximum posterior probability. We have argued that inference in the real world is beset by constraints on available information and computational resources. In this context, optimality must be defined in terms of a tradeoff that balances the accuracy or benefit of the inference against an appropriate information or computational cost. Concretely, we might say that an inference procedure is optimal if it maximizes benefit per unit cost, in which the benefit is some monotonically increasing function of accuracy. An example of this kind of objective is benefit per unit time. In fact, in some cases optimization of this objective is straightforward to implement by adjusting the amount of evidence that is integrated in the decision process (the decision threshold in the DDM and similar models) and can help account for reward-driven decision-making behaviors [66, 67, 61, 62, 68, 70, 108].

We propose a generalization of this approach that decomposes the benefit/cost curve into two components (Fig. 2 and [55]). The first component describes the benefit, perhaps the reward obtained, as a function of accuracy. This function generally increases monotonically with accuracy but can take many different forms that reflect the goals or needs of the decision-maker, the task conditions, and other factors [109, 85]. The second component describes accuracy as a function of costs, which can include the time, memory, and computational resources needed to process information [55, 60]. This function is also generally monotonically increasing (e.g., investing more resources yields more accurate solutions in TAFC tasks, dynamic-state inference, and machine-learning applications). Optimizing benefit per cost in this setting will not generally lead to the solution that maximizes benefit, which is the target of classical optimization approaches. For example, consider scenarios in which benefit increases monotonically with accuracy of the inference procedure, and accuracy increases monotonically with the cost of carrying out the inference procedure. Also recall that a typical inference engine (whether a brain or computer) has a fixed resting-state cost to maintain the computational machinery. Below this threshold cost, accuracy will vanish. Here we analyze several scenarios of this kind.

First consider a setting in which benefit (B) is proportional to accuracy (A) so that $B = \alpha A$, while accuracy is a linear function of cost (C) above a threshold (t) so that $A = \beta(C–t)$, and vanishes when $C < t$. So if $C < t$, the ratio B/C vanishes. But for $C > t$, $B/C = \alpha\beta(C - t)/C = \alpha\beta\,(1 - t/C)$ which is maximized at the largest cost, or equivalently at the highest benefit. Thus, if benefit grows linearly with accuracy, and accuracy grows linearly with cost over a threshold, benefit/cost is maximized by simply maximizing the benefit, as expected in traditional optimization approaches. Next consider a setting (Fig. 2, column A) in which benefit is a convex (super-linear) function of accuracy, while accuracy is a convex function of cost above a threshold. Here doubling the cost more than doubles the accuracy, which in turn more than doubles the benefit. These relationships imply that benefit/cost is again maximized when benefit is maximized, in this case at the highest cost.

More realistically, consider scenarios (Fig. 2, column B) in which benefit is a concave (sub-linear) function of accuracy, while accuracy is a concave function of cost above a threshold. Concavity here implies a law of diminishing returns [55, 101, 110]: doubling the cost over threshold yields less than double the accuracy, which in turn gives less than double the benefit. In this setting, the benefit/cost is maximized at an intermediate cost and thus at less than maximal accuracy and benefit (red dots in Fig. 2, column B). Columns C and D of Fig. 2 demonstrate the same result for a saturating (sigmoidal) benefit-accuracy curve with both sigmoidal and concave accuracy-cost curves. In fact, this tradeoff appears very generally in scenarios in which there is a threshold cost and a subsequent law of diminishing returns, for example in energy-efficient information transmission [110] in which information/energy is maximized at some intermediate information rate.

These results suggest that when animals use strategies that do not maximize accuracy or other benefits, they may actually be rationally trading off benefits against costs that reflect constraints on information and computation.

Acknowledgments

GT is supported by the Swartz Foundation and the Computational Neuroscience Initiative of the University of Pennsylvania. VB and JG are supported in part by NIH BRAIN Initiative grant R01EB026945. JG is also supported by R01 MH115557 and NSF-NCS 1533623.

References

[1]. Geisler WS: Contributions of ideal observer theory to vision research. Vision Research 2011, 51:771–781. [PubMed: 20920517]

[2]. Xu F, Tenenbaum JB: Word learning as Bayesian inference. Psychological Review 2007, 114:245–272. [PubMed: 17500627]

[3]. Perfors A, Navarro DJ: What Bayesian modelling can tell us about statistical learning: what it requires and why it works. Statistical Learning and Language Acquisition 2012, 1:383–408.

[4]. Navarro DJ, Dry MJ, Lee MD: Sampling assumptions in inductive generalization. Cognitive Science 2012, 36:187–223. [PubMed: 22141440]

[5]. Navarro D, Lee M, Dry M, Schultz B: Extending and testing the Bayesian theory of generalization. In Proceedings of the 30th Annual Meeting of the Cognitive Science Society. 2008:1746–1751.

[6]. Kemp C, Perfors A, Tenenbaum JB: Learning overhypotheses with hierarchical Bayesian models. Developmental Science 2007, 10:307–321. [PubMed: 17444972]

[7]. Glassen T, Nitsch V: Hierarchical Bayesian models of cognitive development. Biological Cybernetics 2016, 110:217–227. [PubMed: 27222110]

[8]. Thompson B, de Boer B: Structure and abstraction in phonetic computation: learning to generalise among concurrent acquisition problems. Journal of Language Evolution 2017, 2:94–112.

[9]. Pajak B, Fine AB, Kleinschmidt DF, Jaeger TF: Learning additional languages as hierarchical probabilistic inference: insights from first language processing. Language Learning 2016, 66:900–944. [PubMed: 28348442]

[10]. Pajak B, Bicknell K, Levy R: A model of generalization in distributional learning of phonetic categories. In Proceedings of the Fourth Annual Workshop on Cognitive Modeling and Computational Linguistics (CMCL). 2013:11–20.

[11]. Feldman NH, Griffiths TL, Goldwater S, Morgan JL: A role for the developing lexicon in phonetic category acquisition. Psychological Review 2013, 120:751–778. [PubMed: 24219848]

[12]. Gauthier J, Levy R, Tenenbaum JB: Word learning and the acquisition of syntactic–semantic overhypotheses. 2018 arXiv:1805.04988.

[13]. Navarro DJ, Kemp C: None of the above: a Bayesian account of the detection of novel categories. Psychological Review 2017, 124:643–677. [PubMed: 28703607] (**) A theoretical and experimental investigation on how the statistical structure of observed data affects the speed of category learning.

[14]. Rigoli F, Pezzulo G, Dolan R, Friston K: A goal-directed Bayesian framework for categorization. Frontiers in Psychology 2017, 8:408. [PubMed: 28382008] (*) A Bayesian foundation of hierarchical learning integrating perceptual and contextual aspects in a single framework.

[15]. Gershman SJ, Niv Y: Novelty and inductive generalization in human reinforcement learning. Topics in Cognitive Science 2015, 7:391–415. [PubMed: 25808176]

[16]. Mathys C, Daunizeau J, Friston KJ, Stephan KE: A Bayesian foundation for individual learning under uncertainty. Frontiers in Human Neuroscience 2011, 5:39. [PubMed: 21629826]

[17]. Mathys CD, Lomakina EI, Daunizeau J, Iglesias S, Brodersen KH, Friston KJ, Stephan KE: Uncertainty in perception and the Hierarchical Gaussian Filter. Frontiers in Human Neuroscience 2014, 8:825. [PubMed: 25477800]

[18]. Vossel S, Mathys C, Daunizeau J, Bauer M, Driver J, Friston KJ, Stephan KE: Spatial attention, precision, and Bayesian inference: a study of saccadic response speed. Cerebral Cortex 2013, 24:1436–1450. [PubMed: 23322402]

[19]. Iglesias S, Mathys C, Brodersen KH, Kasper L, Piccirelli M, den Ouden HEM, Stephan KE: Hierarchical prediction errors in midbrain and basal forebrain during sensory learning. Neuron 2013, 80:519–530. [PubMed: 24139048]

[20]. Gershman SJ: On the blessing of abstraction. The Quarterly Journal of Experimental Psychology 2017, 70:361–365. [PubMed: 26930189] (**) A demonstration, through a hierarchical Bayesian model and an experimental study, that abstract knowledge can be acquired faster than specific knowledge when the variance of specific variables conditional on abstract variables is low.

[21]. Adams RP, MacKay DJC: Bayesian online changepoint detection. 2007 arXiv:0710.3742.

[22]. Fearnhead P, Liu Z: On-line inference for multiple changepoint problems. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 2007, 69:589–605.

[23]. Behrens TEJ, Woolrich MW, Walton ME, Rushworth MFS: Learning the value of information in an uncertain world. Nature Neuroscience 2007, 10:1214–1221. [PubMed: 17676057]

[24]. Wilson RC, Nassar MR, Gold JI: Bayesian online learning of the hazard rate in change-point problems. Neural Computation 2010, 22:2452–2476. [PubMed: 20569174]

[25]. Radillo AE, Veliz-Cuba A, Josić K, Kilpatrick ZP: Evidence accumulation and change rate inference in dynamic environments. Neural Computation 2017, 29:1561–1610. [PubMed: 28333591] (*) A novel biologically inspired algorithm for inference of dynamic discrete-valued states and their transition probabilities.

[26]. Veliz-Cuba A, Kilpatrick ZP, Josic K: Stochastic models of evidence accumulation in changing environments. SIAM Review 2016, 58:264–289.

[27]. Glaze CM, Kable JW, Gold JI: Normative evidence accumulation in unpredictable environments. Elife 2015, 4:e08825.

[28]. Deneve S: Bayesian spiking neurons I: Inference. Neural Computation 2008, 20:91–117. [PubMed: 18045002]

[29]. Genewein T, Braun DA: Occam's razor in sensorimotor learning. Proceedings of the Royal Society B: Biological Sciences 2014, 281:20132952.

[30]. Johnson S, Jin A, Keil F: Simplicity and goodness-of-fit in explanation: the case of intuitive curve-fitting. In Proceedings of the Annual Meeting of the Cognitive Science Society. 2014, 36:701–706.

[31]. Roitman JD, Shadlen MN: Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. Journal of Neuroscience 2002, 22:9475–9489. [PubMed: 12417672]

[32]. Kobayashi T, Iwamura M, Matsuda T, Kise K: An anytime algorithm for camera-based character recognition. In 2013 12th International Conference on Document Analysis and Recognition. 2013:1140–1144.

[33]. Balasubramanian V: Statistical inference, Occam's razor, and statistical mechanics on the space of probability distributions. Neural Computation 1997, 9:349–368.

[34]. Myung IJ, Balasubramanian V, Pitt MA: Counting probability distributions: differential geometry and model selection. Proceedings of the National Academy of Sciences 2000, 97:11170–11175.

[35]. Feldman J: The simplicity principle in perception and cognition. Wiley Interdisciplinary Reviews: Cognitive Science 2016, 7:330–340. [PubMed: 27470193]

[36]. Little DRB, Shiffrin R: Simplicity bias in the estimation of causal functions. In Proceedings of the Annual Meeting of the Cognitive Science Society. 2009, 31:1157–1162.

[37]. Zwiernik P, Smith JQ: The dependence of routine Bayesian model selection methods on irrelevant alternatives. 2012 arXiv:1208.3553.

[38]. Busemeyer JR, Wang Z: What is quantum cognition, and how is it applied to psychology? Current Directions in Psychological Science 2015, 24:163–169.

[39]. Bruza PD, Wang Z, Busemeyer JR: Quantum cognition: a new theoretical approach to psychology. Trends in Cognitive Sciences 2015, 19:383–393. [PubMed: 26058709]

[40]. Wang Z, Busemeyer JR: Interference effects of categorization on decision making. Cognition 2016, 150:133–149. [PubMed: 26896726]

[41]. Wojciechowski BW, Pothos EM: Is there a conjunction fallacy in legal probabilistic decision making? Frontiers in Psychology 2018, 9:391. [PubMed: 29674983]

[42]. Khrennikov A: Quantum Bayesianism as the basis of general theory of decision-making. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 2016, 374:20150245.

[43]. Aerts D, de Bianchi MS, Sozzo S, Veloz T: Modeling human decisionmaking: an overview of the Brussels quantum approach. Foundations of Science 2018:1–28.

[44]. Cui X, Ghanem D, Kuffner T: On model selection criteria for climate change impact studies. 2018 arXiv:1808.07861.

[45]. Acuna D, Schrater PR: Structure learning in human sequential decision making. In Advances in Neural Information Processing Systems. 2009:1–8.

[46]. Gigerenzer G, Gaissmaier W: Heuristic decision making. Annual Review of Psychology 2011, 62:451–482.

[47]. Mishra S: Decision-making under risk: integrating perspectives from biology, economics, and psychology. Personality and Social Psychology Review 2014, 18:280–307. [PubMed: 24769798]

[48]. Artinger F, Petersen M, Gigerenzer G, Weibler J: Heuristics as adaptive decision strategies in management. Journal of Organizational Behavior 2015, 36:S33–S52.

[49]. Maitland E, Sammartino A: Decision making and uncertainty: the role of heuristics and experience in assessing a politically hazardous environment. Strategic Management Journal 2015, 36:1554–1578.

[50]. Djulbegovic B, Elqayam S: Many faces of rationality: implications of the great rationality debate for clinical decision-making. Journal of Evaluation in Clinical Practice 2017, 23:915–922. [PubMed: 28730671]

[51]. Marewski JN, Gigerenzer G: Heuristic decision making in medicine. Dialogues in Clinical Neuroscience 2012, 14:77–89. [PubMed: 22577307]

[52]. McLaughlin K, Eva KW, Norman GR: Reexamining our bias against heuristics. Advances in Health Sciences Education 2014, 19:457–464. [PubMed: 24889994]

[53]. Huang L, Pearce JL: Managing the unknowable: the effectiveness of early-stage investor gut feel in entrepreneurial investment decisions. Administrative Science Quarterly 2015, 60:634–670.

[54]. Glaze CM, Filipowicz ALS, Kable JW, Balasubramanian V, Gold JI: A bias–variance trade-off governs individual differences in on-line learning in an unpredictable environment. Nature Human Behaviour 2018, 2:213–224.(**) A demonstration of a bias-variance trade-off in human two-alternative forced-choice inference under noisy and volatile conditions. The trade-off arises from considering broader (more complex) or narrower (simpler) hypothesis spaces over environmental volatility.

[55]. Tavoni G, Balasubramanian V, Gold JI: The complexity dividend: when sophisticated inference matters. 2019 bioRxiv:10.1101/563346.(**) An overarching theory of inference of dynamic latent states from noisy observations under bounded rationality. The theory predicts diminishing returns in accuracy gains with increasing complexity of the computations and that only under a restricted range of conditions complex cognitive operations are required to perform effective inference.

[56]. Lieder F, Gri ths TL: When to use which heuristic: a rational solution to the strategy selection problem. In CogSci. 2015.

[57]. Gershman SJ, Horvitz EJ, Tenenbaum JB: Computational rationality: a converging paradigm for intelligence in brains, minds, and machines. Science 2015, 349:273–278. [PubMed: 26185246]

[58]. Guy TV, Kárnỳ M, Wolpert DH: Decision Making with Imperfect Decision Makers, vol. 28 Springer Science & Business Media; 2011.

[59]. Marcus GF, Davis E: How robust are probabilistic models of higher-level cognition? Psychological Science 2013, 24:2351–2360. [PubMed: 24084039]

[60]. Rahnev D, Denison RN: Suboptimality in perceptual decision making. Behavioral and Brain Sciences 2018, 41:1–66.

[61]. Otto AR, Daw ND: The opportunity cost of time modulates cognitive effort. Neuropsychologia 2019, 123:92–105. [PubMed: 29750987]

[62]. Dambacher M, Hübner R: Time pressure affects the efficiency of perceptual processing in decisions under conflict. Psychological Research 2015, 79:83–94. [PubMed: 24487728]

[63]. Forstmann BU, Ratcli R, Wagenmakers EJ: Sequential sampling models in cognitive neuroscience: advantages, applications, and extensions. Annual Review of Psychology 2016, 67:641–666.

[64]. Shadlen MN, Shohamy D: Decision making and sequential sampling from memory. Neuron 2016, 90:927–939. [PubMed: 27253447]

[65]. Tajima S, Drugowitsch J, Pouget A: Optimal policy for value-based decision-making. Nature Communications 2016, 7:12400.

[66]. Gold JI, Shadlen MN: Banburismus and the brain: decoding the relationship between sensory stimuli, decisions, and reward. Neuron 2002, 36:299–308. [PubMed: 12383783]

[67]. Simen P, Contreras D, Buck C, Hu P, Holmes P, Cohen JD: Reward rate optimization in two-alternative decision making: empirical tests of theoretical predictions. Journal of Experimental Psychology: Human Perception and Performance 2009, 35:1865–1897. [PubMed: 19968441]

[68]. Bogacz R, Brown E, Moehlis J, Holmes P, Cohen JD: The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. Psychological Review 2006, 113:700–765. [PubMed: 17014301]

[69]. Mathias SR, Knowles EEM, Barrett J, Leach O, Buccheri S, Beetham T, Blangero J, Poldrack RA, Glahn DC: The processing-speed impairment in psychosis is more than just accelerated aging. Schizophrenia Bulletin 2017, 43:814–823. [PubMed: 28062652] (*) An interesting use of the drift-diffusion-model to uncover the different influences of psychosis and aging on processing information under time constraints.

[70]. Moran R: Optimal decision making in heterogeneous and biased environments. Psychonomic Bulletin & Review 2015, 22:38–53. [PubMed: 24928091]

[71]. Ortega PA, Braun DA: Thermodynamics as a theory of decision-making with information-processing costs. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences 2013, 469:20120683.

[72]. Anderson ML, Oates T: A review of recent research in metareasoning and metalearning. AI Magazine 2007, 28:7–16.

[73]. Rivera J, Goebel C, Jacobsen HA: A distributed anytime algorithm for real-time EV charging congestion control. In Proceedings of the 2015 ACM Sixth International Conference on Future Energy Systems. 2015:67–76.

[74]. Kil H, Nam W: Efficient anytime algorithm for large-scale QoS-aware web service composition. International Journal of Web and Grid Services 2013, 9:82–106.

[75]. Lin J, Trotman A: Anytime ranking for impact-ordered indexes. In Proceedings of the 2015 International Conference on The Theory of Information Retrieval. 2015:301–304.

[76]. Yang J, Xu Y, Shang Y, Chen G: A space-bounded anytime algorithm for the multiple longest common subsequence problem. IEEE Transactions on Knowledge and Data Engineering 2014, 26:2599–2609. [PubMed: 25400485]

[77]. Koes M, Nourbakhsh I, Sycara K: Heterogeneous multirobot coordination with spatial and temporal constraints. In AAAI. 2005, 5:1292–1297.

[78]. Standley TS, Korf R: Complete algorithms for cooperative pathfinding problems. In Twenty-Second International Joint Conference on Artificial Intelligence. 2011:668–673.

[79]. Macarthur KS, Stranders R, Ramchurn S, Jennings N: A distributed anytime algorithm for dynamic task allocation in multi-agent systems. In Twenty-Fifth AAAI Conference on Artificial Intelligence. 2011:701–706.

[80]. Nerurkar ED, Roumeliotis SI: Power-SLAM: a linear-complexity, anytime algorithm for SLAM. The International Journal of Robotics Research 2011, 30:772–788.

[81]. Karaman S, Walter MR, Perez A, Frazzoli E, Teller S: Anytime motion planning using the RRT. In 2011 IEEE International Conference on Robotics and Automation. 2011:1478–1483.

[82]. Ueno K, Xi X, Keogh E, Lee DJ: Anytime classification using the nearest neighbor algorithm with applications to stream mining. In Sixth International Conference on Data Mining (ICDM’06). 2006:623–632.

[83]. Courville AC, Daw ND: The rat as particle filter. In Advances in Neural Information Processing Systems. 2008:369–376.

[84]. Griffiths TL, Vul E, Sanborn AN: Bridging levels of analysis for probabilistic models of cognition. Current Directions in Psychological Science 2012, 21:263–268.

[85]. Vul E, Goodman N, Griffiths TL, Tenenbaum JB: One and done? Optimal decisions from very few samples. Cognitive Science 2014, 38:599–637. [PubMed: 24467492]

[86]. Smith A: Sequential Monte Carlo Methods in Practice. Springer Science & Business Media; 2013.

[87]. Sanborn AN, Chater N: Bayesian brains without probabilities. Trends in Cognitive Sciences 2016, 20:883–893. [PubMed: 28327290]

[88]. Shi L, Griffiths TL: Neural implementation of hierarchical Bayesian inference by importance sampling. In Advances in Neural Information Processing Systems. 2009:1669–1677.

[89]. Wilson RC, Nassar MR, Gold JI: A mixture of delta-rules approximation to Bayesian inference in change-point problems. PLoS Computational Biology 2013, 9:e1003150. [PubMed: 23935472]

[90]. Wilson RC, Nassar MR, Tavoni G, Gold JI: Correction: A mixture of delta-rules approximation to Bayesian inference in change-point problems. PLoS computational biology 2018, 14:e1006210. [PubMed: 29944654]

[91]. Nassar MR, Wilson RC, Heasly B, Gold JI: An approximately Bayesian delta-rule model explains the dynamics of belief updating in a changing environment. Journal of Neuroscience 2010, 30:12366–12378. [PubMed: 20844132]

[92]. Nassar MR, Rumsey KM, Wilson RC, Parikh K, Heasly B, Gold JI: Rational regulation of learning dynamics by pupil-linked arousal systems. Nature Neuroscience 2012, 15:1040–1046. [PubMed: 22660479]

[93]. Xu X, Ding Y, Hu SX, Niemier M, Cong J, Hu Y, Shi Y: Scaling for edge inference of deep neural networks. Nature Electronics 2018, 1:216–222.(*) A perspective on the gaps existing between Moore's law and the scaling of DNNs for edge inference, and on some architecture and algorithm innovations that could help to bridge these gaps.

[94]. Wong A, Shafiee MJ, Chwyl B, Li F: FermiNets: learning generative machines to generate effcient neural networks via generative synthesis. 2018 arXiv:1809.05989.

[95]. Wong A, Shafiee MJ, Jules MS: MicronNet: a highly compact deep convolutional neural network architecture for real-time embedded traffic sign classification. IEEE Access 2018, 6:59803–59810.

[96]. Abtahi T, Shea C, Kulkarni A, Mohsenin T: Accelerating convolutional neural network with FFT on embedded hardware. IEEE Transactions on Very Large Scale Integration (VLSI) Systems 2018, 26:1737–1749.

[97]. Iandola F, Keutzer K: Small neural nets are beautiful: enabling embedded systems with small deep-neural-network architectures. In Proceedings of the Twelfth IEEE/ACM/IFIP International Conference on Hardware/Software Codesign and System Synthesis Companion. 2017:1–10.

[98]. Tang R, Wang W, Tu Z, Lin J: An experimental analysis of the power consumption of convolutional neural networks for keyword spotting. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2018:5479–5483.

[99]. Tang R, Lin J: Adaptive pruning of neural language models for mobile devices. 2018 arXiv:1809.10282.

[100]. Taylor B, Marco VS, Wol W, Elkhatib Y, Wang Z: Adaptive selection of deep learning models on embedded systems. 2018 arXiv:1805.04252.

[101]. Canziani A, Paszke A, Culurciello E: An analysis of deep neural network models for practical applications. 2016 arXiv:1605.07678.

[102]. Wong A: NetScore: towards universal metrics for large-scale performance analysis of deep neural networks for practical on-device edge usage. 32nd Conference on Neural Information Processing Systems (NIPS 2018) 2018.

[103]. Tishby N, Zaslavsky N: Deep learning and the information bottleneck principle. In 2015 IEEE Information Theory Workshop (ITW). 2015:1–5.

[104]. Schwartz-Ziv R, Tishby N: Opening the black box of deep neural networks via information. 2017 arXiv:1703.00810.(**) An insightful theoretical interpretation of SGD-based training of DNNs as a learning scheme that optimizes the "Information Bottleneck" tradeoff between compression of the input and prediction performance.

[105]. Tishby N, Pereira FC, Bialek W: The information bottleneck method. 2000 arXiv:physics/0004057.

[106]. Huang J, Rathod V, Sun C, Zhu M, Korattikara A, Fathi A, Fischer I, Wojna Z, Song Y, Guadarrama S, Murphy K: Speed/accuracy trade-off for modern convolutional object detectors. In Proceedings of the IEEE conference on computer vision and pattern recognition. 2017:7310–7311.(*) An interesting approach, based on a unified and flexible implementation of different DNN meta-architectures for object detection, to trace out the speedaccuracy trade-off of classes of DNNs.

[107]. Rodrigues CF, Riley G, Luján M: SyNERGY: an energy measurement and prediction framework for convolutional neural networks on Jetson TX1. In Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA). 2018:375–382.

[108]. Fan Y, Gold JI, Ding L: Ongoing, rational calibration of reward-driven perceptual biases. eLife 2018, 7:e36018. [PubMed: 30303484]

[109]. Simon HA: Rational choice and the structure of the environment. Psychological Review 1956, 63:129–138. [PubMed: 13310708]

[110]. Balasubramanian V, Kimber D, Ii MJB: Metabolically efficient information processing. Neural computation 2001, 13:799–815. [PubMed: 11255570]

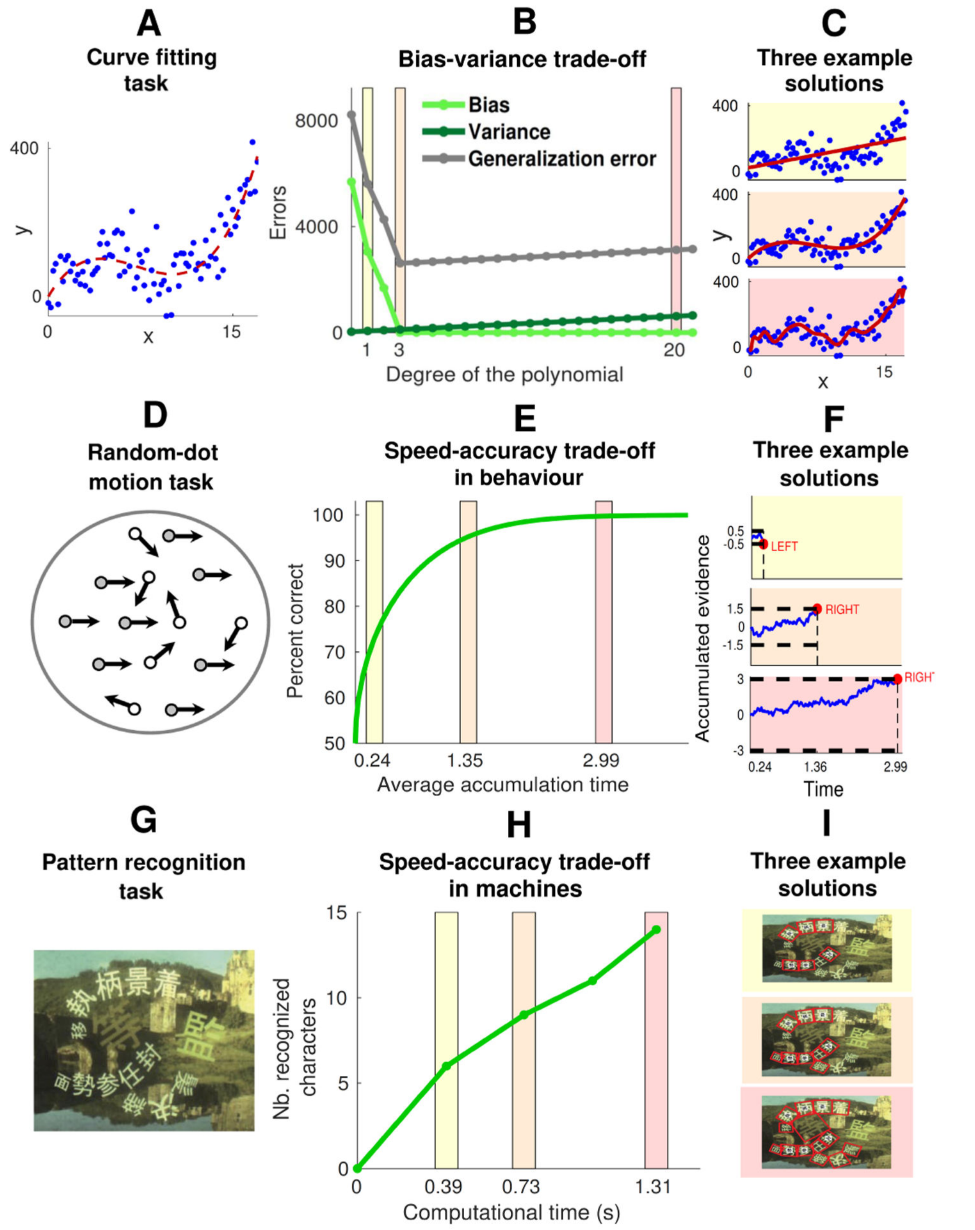

Figure 1: Trade-offs in inference with limited information, time, and computational resources. **Top panels**: The bias-variance trade-off. A A curve-fitting task that requires an inference about the hidden curve (red dashed line) that is most likely to generate the data points (blue dots) with Gaussian noise [29, 30]. **B** Under limited data, increasing the degree of the fitting polynomial (and hence the statistical complexity of the solution) decreases errors due to bias (underfitting) but increases errors due to variance (overfitting). The total generalization error is minimized at intermediate complexity. **C** Three example solutions of increasing statistical complexity (yellow to pink); intermediate is optimal in this case. **Middle panels**: An example of speed-accuracy trade-off in behaviour. **D** A random-dot motion task that requires an inference about the dominant direction of motion of stochastic visual dots [31]. **E** The percentage of correct responses can be increased by increasing the time to sequentially

process information (accumulation time) about the direction of motion of the dots. **F** Example solutions showing increased accuracy but longer decision times as the pre-defined bound on the total evidence to integrate in the decision process (black dashed line) increases (yellow to pink). **Bottom panels:** An example of speed-accuracy trade-off in machines (adapted from [32]). **G** A pattern-recognition task that requires the identification of characters embedded in a scene image. **H** This task can be solved by an “anytime” algorithm that is governed by a trade-off between accuracy and computational time to process information in the image. **I** As the running time increases, the algorithm localizes (red rectangles) more and more characters (yellow to pink).

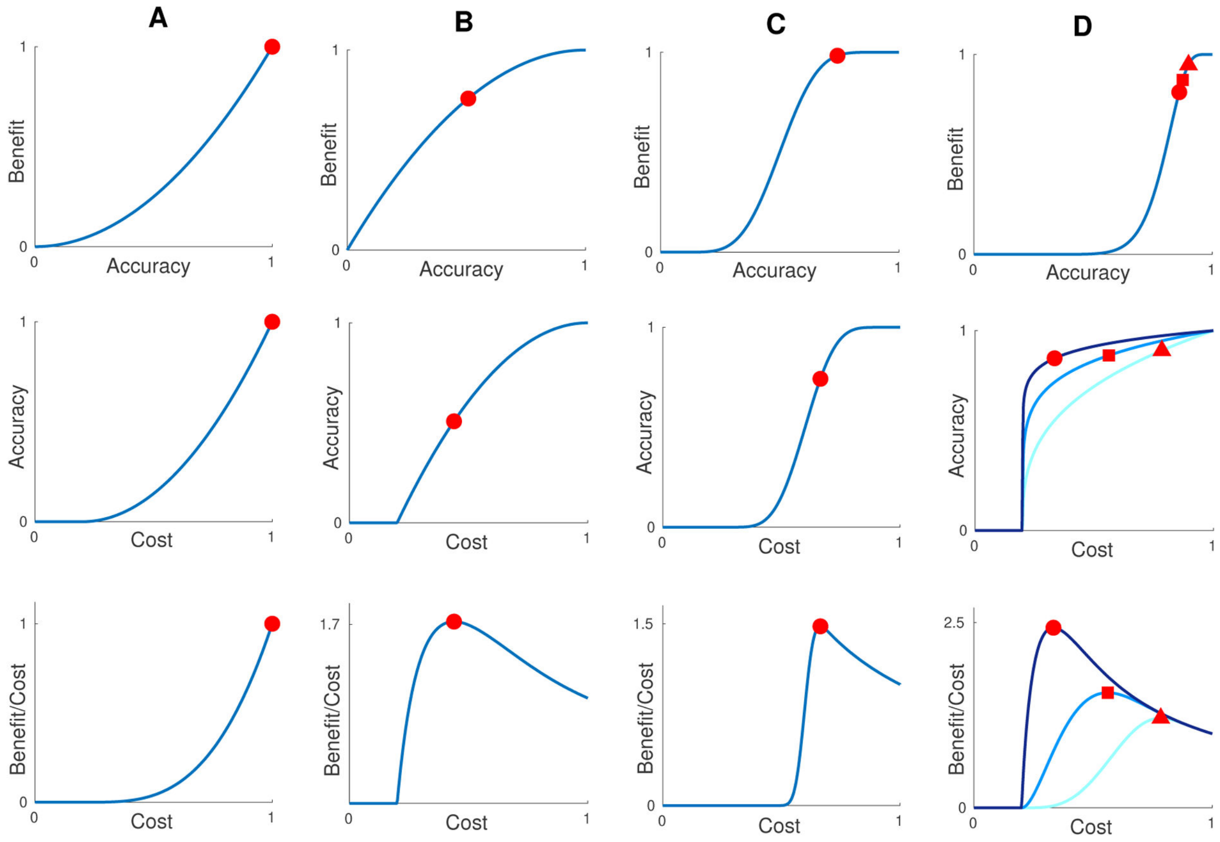

Figure 2: Different Benefit vs. Accuracy (top) and Accuracy vs. Cost (middle) curves yield optimal solutions that vary widely in cost and accuracy.

The optimum is defined as the maximum of the Benefit/Cost ratio (red markers). We consider cases where Benefit increases monotonically as a function of the Accuracy, and Accuracy increases monotonically as a function of Cost. There is minimum operating cost, which means that Accuracy vanishes if Cost is less than this minimum. We consider four general scenarios for the two functions: (**A**) convex, (**B**) concave, (**C**) sigmoid, and (**E**), a common scenario in which there is little Benefit below a threshold Accuracy and maximal Benefit is quickly attained above this threshold, while Accuracy is a concave function of Cost. Concave functions encode a law of diminishing returns – e.g., if Accuracy is a concave function of Cost, then doubling the Cost gives less than double the return in Accuracy. In this scenario, different rates of diminishing returns (different blue lines) give optimal solutions with widely different costs (different red markers) [55].