AFRL-RI-RS-TR-2021-199

HASTY: A GENERATIVE MODEL COMPILER

UNIVERSITY OF BRITISH COLOMBIA

NOVEMBER 2021

FINAL TECHNICAL REPORT

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE

■ **AIR FORCE MATERIEL COMMAND** ■ **UNITED STATES AIR FORCE** ■ **ROME, NY 13441**

NOTICE AND SIGNATURE PAGE

AFRL-RI-RS-TR-2021-199 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

FOR THE CHIEF ENGINEER:

/ S /
MICHAEL J. MANNO
Work Unit Manager

/ S /
SCOTT D. PATRICK
Deputy Chief
Intelligence Systems Division
Information Directorate

REPORT DOCUMENTATION PAGE

1. REPORT DATE NOVEMBER 2021	**2. REPORT TYPE** FINAL TECHNICAL REPORT	**3. DATES COVERED** **START DATE** AUGUST 2019	**END DATE** MAY 2021

4. TITLE AND SUBTITLE

HASTY: A GENERATIVE MODEL COMPILER

5a. CONTRACT NUMBER FA8750-19-2-0222	**5b. GRANT NUMBER** N/A	**5c. PROGRAM ELEMENT NUMBER** 62702E
5d. PROJECT NUMBER D3ME	**5e. TASK NUMBER** 00	**5f. WORK UNIT NUMBER** 14

6. AUTHOR(S)

Frank Wood

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

University of British Colombia
2329 West Mall
Vancouver BC Canada V6T 1Z4

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) Air Force Research Laboratory/RIED 525 Brooks Road Rome NY 13441-4505	**10. SPONSOR/MONITOR'S ACRONYM(S)** RI	**11. SPONSOR/MONITOR'S REPORT NUMBER(S)** AFRL-RI-RS-TR-2021-199

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for Public Release; Distribution Unlimited. This report is the result of contracted fundamental research deemed exempt from public affairs security and policy review in accordance with SAF/AQR memorandum dated 10 Dec 08 and AFRL/CA policy clarification memorandum dated 16 Jan 09.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

DARPA's Data Driven Discovery of Models (D3M) program aimed to automate machine learning so as to allow non-expert government users to pose queries, make predictions, and otherwise model data without necessarily having a background in data science or machine learning. Our research was to develop an extensive set of model primitives and contribute them to a library of discoverable component models that would be assembled by other teams' pipeline search and tuning software tools. Our part of this work was developing a software toolchain that made it possible to "vastly" extend a set of such model primitives. Our contributions included significantly influencing the design of the D3M primitive interface, implementing the original core primitives that demonstrated this interface, implementing an additional number of specialized primitives, and explaining the opportunity, untapped, for program participants to construct their own primitives based on our developed and contributed generative model compiler technology.

15. SUBJECT TERMS

D3M, software toolchain, model compiler technology, daphne, pyprob

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**
a. REPORT U	**b. ABSTRACT** U	**C. THIS PAGE** U	SAR	33

19a. NAME OF RESPONSIBLE PERSON **MICHAEL J. MANNO**	**19b. PHONE NUMBER** *(Include area code)* **N/A**

Table of Contents

List of Figures ii
List of Tables ii
1.0 Summary 1
2.0 Introduction 2
3.0 Methods, Assumptions and Procedures 3
4.0 Results and Discussion 7
4.1 Training of Highly Qualified Personnel 7
4.2 Scientific Outcomes 7
4.3 Software and Datasets Contributed to D3M Codebase 24
5.0 Conclusions 26
6.0 References 27

List of Figures

Figure 1: Overview of PyProb workflow....4
Figure 2: Overview of Ensemble2 workflow....5
Figure 3: Correlation in test set performance between Ensemble2's base AutoML systems....13

List of Tables

Table 1: Base AutoML Systems used in Ensemble^2....6

1.0 Summary

DARPA’s Data Driven Discovery of Models (D3M) program aimed to automate machine learning so as to allow non-expert government users to pose queries, make predictions, and otherwise model data without necessarily having a background in data science or machine learning. UBC’s proposed role in this large, joint endeavor, was to develop an extensive set of model primitives and contribute them to a library of discoverable component models that would be assembled by other teams’ pipeline search and tuning software tools.

In this context, we were tasked with developing a software toolchain that made it possible to “vastly” extend a set of such model primitives. Due to practical aspects related to the organization of the program, our actual contributions, beyond original science captured in the publication list in this report, included significantly influencing the design of the D3M primitive interface, implementing the original core primitives that demonstrated this interface, implementing an additional number of specialized primitives, and explaining the opportunity, untapped, for program participants to construct their own primitives based on our developed and contributed generative model compiler technology, namely Daphne and PyProb probabilistic programming compilers. This report details these contributions, and some of the fundamental research performed towards these contributions.

2.0 Introduction

The aim of the DARPA Data Driven Discovery of Models (D3M) program was to automate machine learning, removing the need for a data scientist or machine learning expert to perform model development for routine tasks. As a use case, a non-expert government user can pose queries, make predictions, and otherwise model data using the system without a background in data science or machine learning. The core idea in automated machine learning (AutoML) is to algorithmically construct end-to-end machine learning pipelines from a set of model primitives that are automatically assembled and optimized for each new problem at hand.

Our proposed role in this large, joint endeavor, was to develop an extensive set of model primitives and contribute them to a library of discoverable component models that would be assembled by other teams' pipeline search and tuning software tools. Our contributions fit within the context of Technical Activity 1 (TA1) within the program.

Our main goal was developing a software toolchain that made it possible to "vastly" extend a set of such model primitives. Due to practical aspects related to the organization of the program, our actual contributions, beyond original science captured in the publication list below, included significantly influencing the design of the D3M primitive interface,[1] implementing the original core primitives that demonstrated this interface,[2] implementing an additional number of specialized primitives,[3] and explaining the opportunity, untapped, for program participants to construct their own primitives based on our developed and contributed generative model compiler technology.

The software toolchain we developed for vastly expanding the set of model primitives was based on a compiler technology that makes bottom-up, discriminative models from top-down, generative models specified in the form of probabilistic programs. We proposed to call this compiler and runtime Hasty. For various reasons this compiler ended up being two compilers, one named Daphne[4] [1] and another named PyProb[5] [2-4].

In section 3, we will highlight the main principles behind our software toolchain to vastly expand the set of model primitives. Section 4 outlines human resource, scientific, and software outcomes of our work, alongside some details of publications resulting from our contributions.

[1] https://gitlab.com/datadrivendiscovery/d3m/-/tree/devel/d3m/primitive_interfaces
[2] https://gitlab.com/datadrivendiscovery/common-primitives
[3] https://gitlab.com/datadrivendiscovery/primitives/-/tree/master/primitives/UBC
[4] https://github.com/plai-group/daphne
[5] https://github.com/pyprob/pyprob

3.0 Methods, Assumptions and Procedures

Daphne and PyProb transform probabilistic program source code, i.e. code that denotes a generative model and an inference task, into target code for a run-time that makes inference in this generative model efficient via feed forward computation in a trained, bottom-up discriminative model. The target code produced by these systems include two major parts: (1) code to construct a neural network architecture that is dependent on the semantics of the original generative model in the sense that the network will be trained to perform a kind of approximate inference and (2) values for the weights of this neural network. As a computationally intensive compile-time step, these neural network weights are trained so as to make the neural network perform an efficient feed-forward computation that enables extremely efficient approximate posterior inference in the model specified by the original probabilistic program, turning every new compiled probabilistic program, effectively, into a bottom-up discriminative model that can be added to a library of model primitives.

To explain by way of an example, imagine an unstructured record like a telephone number string. These come in a variety of formats such as +12028492838, +44 (0) 7958 123 131, 443-5311, (212)646- 7177, etc. There is a dependently typed structured representation for all these formats: country code, local or area code, and number. One can relatively easily write a generative model that goes from the structured representation to an unstructured, noisy observation. The posterior distribution of structure representations given (conditioned on) an observed telephone number is a model-based primitive for telephone number feature extraction. An amortized inference artifact for such a model is a bespoke custom primitive that could be integrated into an AutoML system for automatic structured featurization of string fields that have the putative type of telephone number. And this kind of primitive can be produced from just a generative model specification. Arguably, in this instance, some very carefully specified regular expression could do a job like the described learned primitive, it almost certainly would be more brittle in the sense of not working under noise distribution shift.

We actually produced this primitive and other example primitives like it (unstructured culture-sensitive approximate name structured interpretable featurization, unstructured date/time string structured interpretable featurization, etc.).

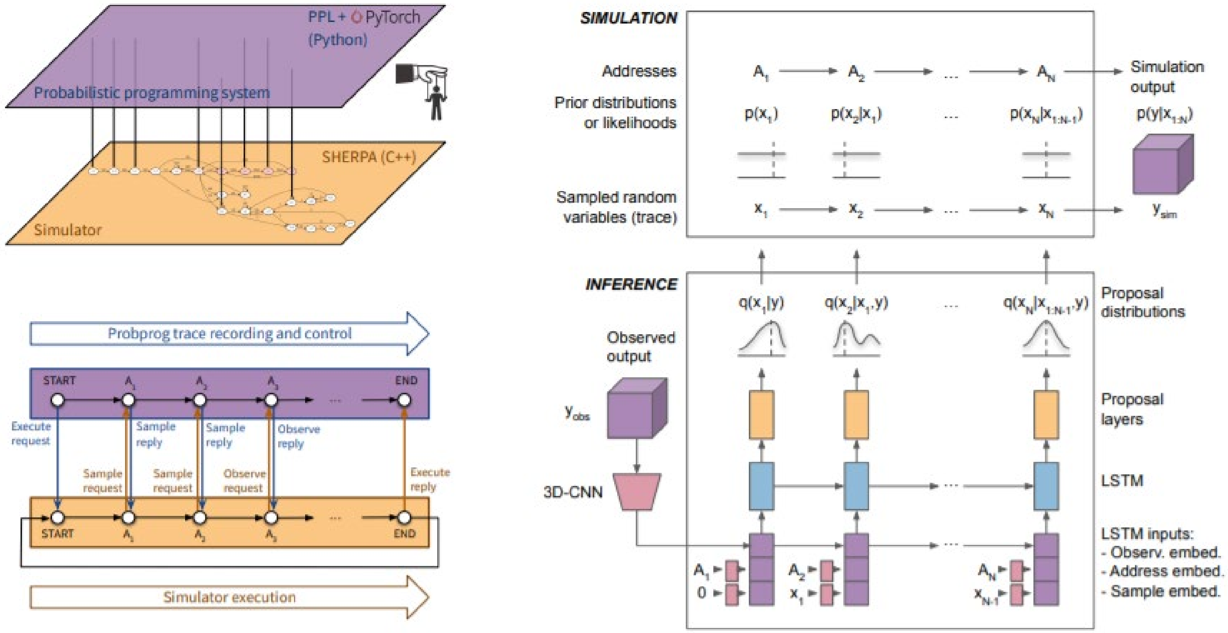

Figure 1. Overview of PyProb workflow [4]. Top Left: The probabilistic execution protocol (PPX). Sample and observe statements correspond to random number draws and conditioning, respectively. Bottom Left: probabilistic execution of a single trace. Top Right: Simulation and inference model addresses, priors and samples. Bottom Right: IC inference engine proposals and NN architecture.

That being said, we did develop two different generative model amortized inference compilers. Daphne in particular compiles first order probabilistic programs to graphical models, automatically inverts their dependency structure, and even goes so far as to automatically specify and train the structure of a flow-based variational family that performs efficient amortized inference in the original generative model. Far more heavyweight, PyProb is an advanced probabilistic programming language that has facilities for automatically inverting stochastic simulators, as outlined in Figure 1. These two compilers fulfill our deliverable promises under our proposal, and, were the program to have run longer, a large number of primitives, particularly of the featurization kind, would have been generated by these tools according to various "customer" specifications for the kind of data and problems they had.

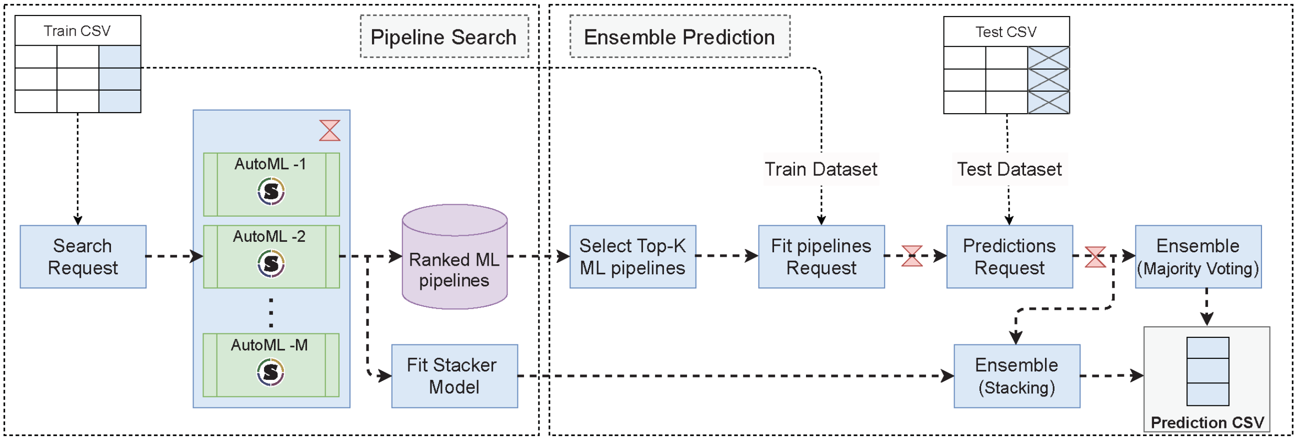

Figure 2. Overview of Ensemble2 workflow. Ensemble2 is made up of two underlying subsystems. The training dataset (in CSV format), along with a user-specified target column, is sent to the Pipeline Search Subsystem. This spins up M different AutoML systems (our base systems) as Singularity containers and performs the pipeline search procedure in parallel for a set time duration. Once the time limit has been hit and all discovered pipelines (P) have been collected, it ranks the pipelines based on their validation scores. The ranked pipelines, alongside the training and test datasets are then passed on to the second subsystem, the Ensemble Prediction Subsystem. Here, the top K pipelines (where K is an Ensemble2 hyperparameter), are selected, optionally refit, and made to generate predictions on the test dataset. The predictions are then passed onto the ensembling module, which generates the final Ensemble2 predictions either using majority voting or a stacking model. The stacking model can choose to ensemble the best pipeline from each AutoML system as well.

Table 1. Base AutoML Systems used in Ensemble^2. This table highlights some of the differences between these AutoML systems and the diversity of methods that Ensemble2 benefits from.

AutoML System	Primitive Library	Model Discovery and Hyperparameter Tuning	Internal Ensembling
AutoGluon	Gluon Library	Fixed Defaults	Multi-Layer Stacking and Bagging
Auto-SkLearn	Scikit-Learn	BayesOpt + Meta-Learning	Forward Search
Auto-SkLearn 2.0	Scikit-Learn	Portfolio Learning	Forward Search
CMU AutoML	D3M Primitives	Templates + Grid Search	-
H2O AutoML	H2O Library	Grid and Random Search	Super Learner

In service of the problem, we made one final but critical contribution to the program. We developed an end-to-end AutoML system of our own called Ensemble^2 [5] along with a web user interface[6] that ensembles D3M program participant AutoML systems together with modern commercial AutoML systems existing at the time of its development. As outlined in Figure 2, this system works by making use of the diversity in the primitives and search process of its underlying AutoML systems, outlined in Table 1. It achieved the highest tabular classification AutoML results in the world at the time of its writing, and provided a means for us to compare and benchmark different AutoML systems against D3M systems, both in terms of pipeline search methods and primitives.

[6] https://blackboxml.cs.ubc.ca/

4.0 Results and Discussion

We have broken down the results of our work into three sections: training of highly qualified personnel with knowledge of the D3M program, scientific outcomes in the form of scholarly publication, and software and datasets.

4.1 Training of Highly Qualified Personnel

The funding from DARPA D3M was used to support 3 principal investigators (Frank Wood, Kevin Leyton-Brown and Katrina Liggett). During the course of the program, the PIs supported over 16 PhD students, 5 Masters students and 2 postdocs using the funding provided by DARPA, alongside a software engineer and program manager to support contributions to the D3M program. The identities of these sponsored HQP is apparent from the author lists of the papers listed in Section 4.2.

4.2 Scientific Outcomes

The work supported by DARPA D3M led to 41 scholarly publications. In the subsections below, we have organized these papers into D3M related subject areas.

4.2.1 Primitives.

Some of the work funded by the program pertained to the development of novel primitives outside the bounds of the inference compilation approach outlined in the original proposal. Notable among these primitives was the SimpleCNAPS primitive [6, 7], which achieved 6.1% improvement on state of the art in few-shot image classification. These include:

Harvey et al.: Image Completion via Inference in Deep Generative Models

W. Harvey, S. Naderiparizi, and F. Wood. “Image Completion via Inference in Deep Generative Models”. 2021. URL: https://arxiv.org/abs/2102.12037.

Abstract: We consider image completion from the perspective of amortized inference in an image generative model. We leverage recent state of the art variational auto-encoder architectures that have been shown to produce photo-realistic natural images at non-trivial resolutions. Through amortized inference in such a model we can train neural artifacts that produce diverse, realistic image completions even when the vast majority of an image is missing. We demonstrate superior sample quality and diversity compared to prior art on the CIFAR-10 and FFHQ-256 datasets. We conclude by describing and demonstrating an application that requires an in-painting model with the capabilities ours exhibits: the use of Bayesian optimal experimental design to select the most informative sequence of small field of view x-rays for chest pathology detection.

Bateni et al.: Improving Few-Shot Visual Classification with Unlabelled Examples

P. Bateni, J. Barber, J.-W. van de Meent, and F. Wood. “Improving Few-Shot Visual Classification with Unlabelled Examples”. 2020. URL: https://arxiv.org/abs/2006.12245.

Abstract: We propose a transductive meta-learning method that uses unlabelled instances to improve few-shot image classification performance. Our approach combines a regularized Mahalanobis-distance-based soft k-means clustering procedure with a state of the art neural adaptive feature extractor to achieve improved test-time classification accuracy using unlabelled data. We evaluate our method on transductive few-shot learning tasks, in which the goal is to jointly predict labels for query (test) examples given a set of support (training) examples. We achieve new state of the art in-domain performance on Meta-Dataset, and improve accuracy on mini- and tiered-ImageNet as compared to other conditional neural adaptive methods that use the same pre-trained feature extractor.

Bateni et al.: Improved Few-Shot Visual Classification
P. Bateni, R. Goyal, V. Masrani, F. Wood, and L. Sigal. “Improved Few-Shot Visual Classification”. In: Conference on Computer Vision and Pattern Recognition (CVPR). 2020. arXiv: 1912.03432. URL: https://arxiv.org/abs/1912.03432.

Abstract: Few-shot learning is a fundamental task in computer vision that carries the promise of alleviating the need for exhaustively labeled data. Most few-shot learning approaches to date have focused on progressively more complex neural feature extractors and classifier adaptation strategies, as well as the refinement of the task definition itself. In this paper, we explore the hypothesis that a simple class-covariance-based distance metric, namely the Mahalanobis distance, adopted into a state of the art few-shot learning approach (CNAPS) can, in and of itself, lead to a significant performance improvement. We also discover that it is possible to learn adaptive feature extractors that allow useful estimation of the high dimensional feature covariances required by this metric from surprisingly few samples. The result of our work is a new "Simple CNAPS" architecture which has up to 9.2% fewer trainable parameters than CNAPS and performs up to 6.1% better than state of the art on the standard few-shot image classification benchmark dataset.

4.2.2 Inference

Some of our funded research focused on foundational inference algorithms, particularly of the form of various approaches to efficiently learning amortized inference artifacts. These contributions included:

Brekelmans et al.: Annealed Importance Sampling with q-Paths
R. Brekelmans, V. Masrani, T. Bui, F. Wood, A. Galstyan, G. V. Steeg, and F. Nielsen. “Annealed Importance Sampling with q-Paths”. 2020. URL: https://arxiv.org/abs/2012.07823.

Abstract: Annealed importance sampling (AIS) is the gold standard for estimating partition functions or marginal likelihoods, corresponding to importance sampling over a path of distributions between a tractable base and an unnormalized target. While AIS yields an unbiased estimator for any path, existing literature has been primarily limited to the geometric mixture or moment-averaged paths associated with the exponential family and KL divergence. We explore AIS using q-paths, which include the geometric path as a special case and are related to the homogeneous power mean, deformed exponential family, and α-divergence.

Nguyen et al.: Gaussian Process Bandit Optimization of the Thermodynamic Variational Objective
V. Nguyen, V. Masrani, R. Brekelmans, M. Osborne, and F. Wood. “Gaussian Process Bandit Optimization of the Thermodynamic Variational Objective”. In: Advances in Neural Information Processing Systems (NeurIPS). 2020. URL: https://arxiv.org/abs/2010.15750.

Abstract: Achieving the full promise of the Thermodynamic Variational Objective (TVO), a recently proposed variational lower bound on the log evidence involving a one-dimensional Riemann integral approximation, requires choosing a "schedule" of sorted discretization points. This paper introduces a bespoke Gaussian process bandit optimization method for automatically choosing these points. Our approach not only automates their one-time selection, but also dynamically adapts their positions over the course of optimization, leading to improved model learning and inference. We provide theoretical guarantees that our bandit optimization converges to the regret-minimizing choice of integration points. Empirical validation of our algorithm is provided in terms of improved learning and inference in Variational Autoencoders and Sigmoid Belief Networks.

Rainforth et al.: Target-Aware Bayesian Inference: How to Beat Optimal Conventional Estimators

T. Rainforth, A. Golinski, F. Wood, and S. Zaidi. "Target-Aware Bayesian Inference: How to Beat Optimal Conventional Estimators". In: Journal of Machine Learning Research 21.88 (2020), pp. 1–54. URL: http://jmlr.org/papers/ v21/19-102.html.

Abstract: Standard approaches for Bayesian inference focus solely on approximating the posterior distribution. Typically, this approximation is, in turn, used to calculate expectations for one or more target functions—a computational pipeline that is inefficient when the target function(s) are known upfront. We address this inefficiency by introducing a framework for target–aware Bayesian inference (TABI) that estimates these expectations directly. While conventional Monte Carlo estimators have a fundamental limit on the error they can achieve for a given sample size, our TABI framework is able to breach this limit; it can theoretically produce arbitrarily accurate estimators using only three samples, while we show empirically that it can also breach this limit in practice. We utilize our TABI framework by combining it with adaptive importance sampling approaches and show both theoretically and empirically that the resulting estimators are capable of converging faster than the standard O(1/N) Monte Carlo rate, potentially producing rates as fast as O(1/N2). We further combine our TABI framework with amortized inference methods, to produce a method for amortizing the cost of calculating expectations. Finally, we show how TABI can be used to convert any marginal likelihood estimator into a target aware inference scheme and demonstrate the substantial benefits this can yield.

Naderiparizi et al.: Uncertainty in Neural Processes

S. Naderiparizi, K. Chiu, B. Bloem-Reddy, and F. Wood. "Uncertainty in Neural Processes". 2020. URL: https://arxiv.org/abs/2010.03753.

Abstract: We explore the effects of architecture and training objective choice on amortized posterior predictive inference in probabilistic conditional generative models. We aim this work to be a counterpoint to a recent trend in the literature that stresses achieving good samples when the amount of conditioning data is large. We instead focus our attention on the case where the amount of conditioning data is small. We highlight specific architecture and objective choices that we find lead to qualitative and quantitative improvement to posterior inference in this low data regime. Specifically we explore the effects of choices of pooling operator and variational family on posterior quality in neural processes. Superior posterior predictive samples drawn from our novel neural process architectures are demonstrated via image completion/in-painting experiments.

Naderiparizi et al.: Amortized rejection sampling in universal probabilistic programming

S. Naderiparizi, A. S´cibior, A. Munk, M. Ghadiri, A. G. Baydin, B. Gram-Hansen, C. S. de Witt, R. Zinkov, P. H. Torr, T. Rainforth, et al. "Amortized rejection sampling in universal probabilistic programming". In: International Conference on Probabilistic Programming (PROBPROG). 2020. arXiv: 1910.09056. URL: https://arxiv.org/abs/1910.09056.

Abstract: Existing approaches to amortized inference in probabilistic programs with unbounded loops can produce estimators with infinite variance. An instance of this is importance sampling inference in programs that explicitly include rejection sampling as part of the user-programmed generative procedure. In this paper we develop a new and efficient amortized importance sampling estimator. We prove finite variance of our estimator and empirically demonstrate our method's correctness and efficiency compared to existing alternatives on generative programs containing rejection sampling loops and discuss how to implement our method in a generic probabilistic programming framework.

Brekelmans et al.: All in the Exponential Family: Bregman Duality in Thermodynamic Variational Inference

R. Brekelmans, V. Masrani, F. Wood, G. Ver Steeg, and A. Galstyan. "All in the Exponential Family: Bregman Duality in Thermodynamic Variational Inference". In: Thirty-seventh International Conference on Machine Learning (ICML 2020). July 2020. arXiv: 2007.00642. URL: https://arxiv.org/abs/2007.00642.

Abstract: The recently proposed Thermodynamic Variational Objective (TVO) leverages thermodynamic integration to provide a family of variational inference objectives, which both tighten and generalize the ubiquitous Evidence Lower Bound (ELBO). However, the tightness of TVO bounds was not previously known, an expensive grid search was used to choose a "schedule" of intermediate distributions, and model learning suffered with ostensibly tighter bounds. In this work, we propose an exponential family interpretation of the geometric mixture curve underlying the TVO and various path sampling methods, which allows us to characterize the gap in TVO likelihood bounds as a sum of KL divergences. We propose to choose intermediate distributions using equal spacing in the moment parameters of our exponential family, which matches grid search performance and allows the schedule to adaptively update over the course of training. Finally, we derive a doubly reparameterized gradient estimator which improves model learning and allows the TVO to benefit from more refined bounds. To further contextualize our contributions, we provide a unified framework for understanding thermodynamic integration and the TVO using Taylor series remainders.

Warrington et al.: Coping With Simulators That Don't Always Return

A. Warrington, S. Naderiparizi, and F. Wood. "Coping With Simulators That Don't Always Return". In: The 23rd International Conference on Artificial Intelligence and Statistics (AISTATS). PMLR 108:1748-1758. 2020. arXiv: 1906.05462. URL: http://proceedings.mlr.press/v108/warrington20a.html.

Abstract: Deterministic models are approximations of reality that are easy to interpret and often easier to build than stochastic alternatives. Unfortunately, as nature is capricious, observational data can never be fully explained by deterministic models in practice. Observation and process noise need to be added to adapt deterministic models to behave stochastically, such that they are capable of explaining and extrapolating from noisy data. We investigate and address computational inefficiencies that arise from adding process noise to deterministic simulators that fail to return for certain inputs; a property we describe as "brittle." We show how to train a conditional normalizing flow to propose perturbations such that the simulator succeeds with high probability, increasing computational efficiency.

Harvey et al.: Attention for Inference Compilation

W. Harvey, A. Munk, A. Baydin, A. Bergholm, and F. Wood. “Attention for Inference Compilation”. In: The second International Conference on Probabilistic Programming (PROBPROG). 2020. arXiv: 1910.11961. URL: https://arxiv.org/abs/1910.11961.

Abstract: We present a new approach to automatic amortized inference in universal probabilistic programs which improves performance compared to current methods. Our approach is a variation of inference compilation (IC) which leverages deep neural networks to approximate a posterior distribution over latent variables in a probabilistic program. A challenge with existing IC network architectures is that they can fail to model long-range dependencies between latent variables. To address this, we introduce an attention mechanism that attends to the most salient variables previously sampled in the execution of a probabilistic program. We demonstrate that the addition of attention allows the proposal distributions to better match the true posterior, enhancing inference about latent variables in simulators.

Campbell et al.: Sparse Variational Inference: Bayesian Coresets from Scratch

T. Campbell and B. Beronov. “Sparse Variational Inference: Bayesian Coresets from Scratch”. In: Conference on Neural Information Processing Systems (NeurIPS). 1st prize, Student poster competition, AICan (Annual Meeting, Pan-Canadian AI Strategy, Canadian Institute for Advanced Research). Vancouver, Canada, Dec. 9, 2019. 2019, pp. 11457–11468. arXiv: 1906.03329. URL: https://arxiv.org/abs/1906.03329.

Abstract: The proliferation of automated inference algorithms in Bayesian statistics has provided practitioners newfound access to fast, reproducible data analysis and powerful statistical models. Designing automated methods that are also both computationally scalable and theoretically sound, however, remains a significant challenge. Recent work on Bayesian coresets takes the approach of compressing the dataset before running a standard inference algorithm, providing both scalability and guarantees on posterior approximation error. But the automation of past coreset methods is limited because they depend on the availability of a reasonable coarse posterior approximation, which is difficult to specify in practice. In the present work we remove this requirement by formulating coreset construction as sparsity-constrained variational inference within an exponential family. This perspective leads to a novel construction via greedy optimization, and also provides a unifying information-geometric view of present and past methods. The proposed Riemannian coreset construction algorithm is fully automated, requiring no problem-specific inputs aside from the probabilistic model and dataset. In addition to being significantly easier to use than past methods, experiments demonstrate that past coreset constructions are fundamentally limited by the fixed coarse posterior approximation; in contrast, the proposed algorithm is able to continually improve the coreset, providing state-of-the-art Bayesian dataset summarization with orders-of-magnitude reduction in KL divergence to the exact posterior.

Gram-Hansen et al.: Efficient Bayesian Inference for Nested Simulators
B. Gram-Hansen, C. Schroeder de Witt, R. Zinkov, S. Naderiparizi, A. Scibior, A. Munk, F. Wood, M. Ghadiri, P. Torr, Y. Whye Teh, A. Gunes Baydin, and T. Rainforth. "Efficient Bayesian Inference for Nested Simulators". In: 2nd Symposium on Advances in Approximate Bayesian Inference (AABI). 2019. URL: https://openreview.net/forum?id=rJeMcy2EtH.

Abstract: We introduce two approaches for conducting efficient Bayesian inference in stochastic simulators containing nested stochastic sub-procedures, i.e., internal procedures for which the density cannot be calculated directly such as rejection sampling loops. The resulting class of simulators are used extensively throughout the sciences and can be interpreted as probabilistic generative models. However, drawing inferences from them poses a substantial challenge due to the inability to evaluate even their unnormalised density, preventing the use of many standard inference procedures like Markov Chain Monte Carlo (MCMC). To address this, we introduce inference algorithms based on a two-step approach that first approximates the conditional densities of the individual sub-procedures, before using these approximations to run MCMC methods on the full program. Because the sub-procedures can be dealt with separately and are lower- dimensional than that of the overall problem, this two-step process allows them to be isolated and thus be tractably dealt with, without placing restrictions on the overall dimensionality of the problem. We demonstrate the utility of our approach on a simple, artificially constructed simulator.

Goliński et al.: Amortized Monte Carlo Integration
A. Golinski, F. Wood, and T. Rainforth. "Amortized Monte Carlo Integration". In: Proceedings of the International Conference on Machine Learning (ICML). 2019. arXiv: 1907.08082. URL: https://arxiv.org/pdf/1907.08082. pdf.

Abstract: Current approaches to amortizing Bayesian inference focus solely on approximating the posterior distribution. Typically, this approximation is, in turn, used to calculate expectations for one or more target functions - a computational pipeline which is inefficient when the target function(s) are known upfront. In this paper, we address this inefficiency by introducing AMCI, a method for amortizing Monte Carlo integration directly. AMCI operates similarly to amortized inference but produces three distinct amortized proposals, each tailored to a different component of the overall expectation calculation. At runtime, samples are produced separately from each amortized proposal, before being combined to an overall estimate of the expectation. We show that while existing approaches are fundamentally limited in the level of accuracy they can achieve, AMCI can theoretically produce arbitrarily small errors for any integrable target function using only a single sample from each proposal at runtime. We further show that it is able to empirically outperform the theoretically optimal self-normalized importance sampler on a number of example problems. Furthermore, AMCI allows not only for amortizing over datasets but also amortizing over target functions.

4.2.3 AutoML

We did work directly on aspects of AutoML, including generative model learning, theory, and probabilistic programming systems that provision model learning and amortized inference for machine learning primitive development.

The most important outcome in this area was Ensemble^2:

Yoo et al.: Ensemble Squared: A Meta AutoML System
J. Yoo, T. Joseph, D. Yung, S. A. Nasseri, and F. Wood. “Ensemble Squared: A Meta AutoML System”. 2020. URL: https://arxiv.org/abs/2012.05390.

Abstract: The continuing rise in the number of problems amenable to machine learning solutions, coupled with simultaneous growth in both computing power and variety of machine learning techniques has led to an explosion of interest in automated machine learning (AutoML). This paper presents Ensemble Squared (Ensemble2), a "meta" AutoML system that ensembles at the level of AutoML systems. Ensemble2 exploits the diversity of existing, competing AutoML systems by ensembling the top-performing models simultaneously generated by a set of them. Our work shows that diversity in AutoML systems is sufficient to justify ensembling at the AutoML system level. In demonstrating this, we also establish a new state of the art AutoML result on the OpenML classification challenge.

Ensemble^2 attempts to exploit the diversity in pipeline search space and heuristic within existing AutoML systems including the D3M systems (NYU, CMU, and TAMU), and external systems (Amazon AutoGluon, AutoSklearn-1, AutoSklearn-2, and H2O AutoML). In addition to exploring the benefits of ensembling to improve performance on AutoML tasks, we used this as an opportunity to develop a platform on which we could test and demonstrate the utility of Hasty primitives, identify gaps in D3M problem types by exposing Ensemble2 to the public and harvesting datasets that can benefit from broader suite primitives, and benchmarking D3M systems on the OpenML AutoML Classification Benchmark [8]. We noticed that no AutoML system consistently outperformed all others on all problems. Performance correlation between AutoML systems across the classification benchmark was around 0.8 0.9 as outlined in Figure 3.

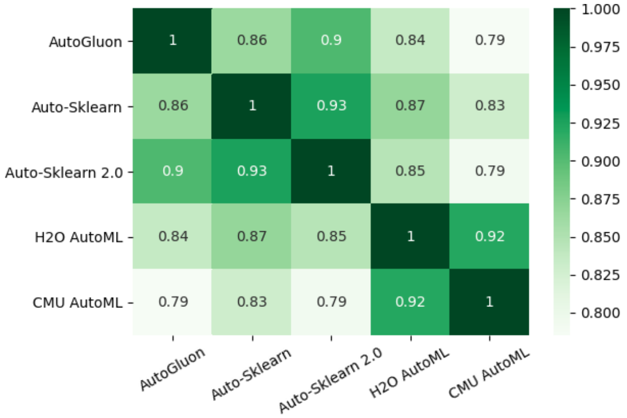

Figure 3. Correlation in test set performance between Ensemble^2's base AutoML systems.

Our work on generative model learning includes:

Le et al.: Revisiting Reweighted Wake-Sleep for Models with Stochastic Control Flow
T. A. Le, A. R. Kosiorek, N. Siddharth, Y. W. Teh, and F. Wood. “Revisiting Reweighted Wake-Sleep for Models with Stochastic Control Flow”. In: ed. by R. P. Adams and V. Gogate. Vol. 115. Proceedings of Machine Learning Research. Tel Aviv, Israel: PMLR, 22–25 Jul 2020, pp. 1039–1049. URL: http://proceedings.mlr.press/v115/le20a. html.

Abstract: Stochastic control-flow models (SCFMs) are a class of generative models that involve branching on choices from discrete random variables. Amortized gradient-based learning of SCFMs is challenging as most approaches targeting discrete variables rely on their continuous relaxations—which can be intractable in SCFMs, as branching on relaxations requires evaluating all (exponentially many) branching paths. Tractable alternatives mainly combine REINFORCE with complex control-variate schemes to improve the variance of naive estimators. Here, we revisit the reweighted wake-sleep (RWS) [5] algorithm, and through extensive evaluations, show that it outperforms current state-of-the-art methods in learning SCFMs. Further, in contrast to the importance weighted autoencoder, we observe that RWS learns better models and inference networks with increasing numbers of particles. Our results suggest that RWS is a competitive, often preferable, alternative for learning SCFMs.

Munk et al.: Assisting the Adversary to Improve GAN Training

A. Munk, W. Harvey, and F. Wood. "Assisting the Adversary to Improve GAN Training". 2020. URL: https://arxiv.org/abs/2010.01274.

Abstract: We propose a method for improved training of generative adversarial networks (GANs). Some of the most popular methods for improving the stability and performance of GANs involve constraining or regularizing the discriminator. Our method, on the other hand, involves regularizing the generator. It can be used alongside existing approaches to GAN training and is simple and straightforward to implement. Our method is motivated by a common mismatch between theoretical analysis and practice: analysis often assumes that the discriminator reaches its optimum on each iteration. In practice, this is essentially never true, often leading to poor gradient estimates for the generator. To address this, we introduce the Adversary's Assistant (AdvAs). It is a theoretically motivated penalty imposed on the generator based on the norm of the gradients used to train the discriminator. This encourages the generator to move towards points where the discriminator is optimal. We demonstrate the effect of applying AdvAs to several GAN objectives, datasets and network architectures. The results indicate a reduction in the mismatch between theory and practice and that AdvAs can lead to improvement of GAN training, as measured by FID scores.

Teng et al.: Semi-supervised Sequential Generative Models

M. Teng, T. A. Le, A. Scibior, and F. Wood. "Semi-supervised Sequential Generative Models". In: Conference on Uncertainty in Artificial Intelligence (UAI). 2020. arXiv: 2007.00155. URL: http://www.auai.org/uai2020/ proceedings/272_main_paper.pdf.

Abstract: We introduce a novel objective for training deep generative time-series models with discrete latent variables for which supervision is only sparsely available. This instance of semi-supervised learning is challenging for existing methods, because the exponential number of possible discrete latent configurations results in high variance gradient estimators. We first overcome this problem by extending the standard semi-supervised generative modeling objective with reweighted wake-sleep. However, we find that this approach still suffers when the frequency of available labels varies between training sequences. Finally, we introduce a unified objective inspired by teacher-forcing and show that this approach is robust to variable length supervision. We call the resulting method caffeinated wake-sleep (CWS) to emphasize its additional dependence on real data. We demonstrate its effectiveness with experiments on MNIST, handwriting, and fruit fly trajectory data.

Masrani et al.: The Thermodynamic Variational Objective
V. Masrani, T. A. Le, and F. Wood. “The Thermodynamic Variational Objective”. In: Thirty-third Conference on Neural Information Processing Systems (NeurIPS). 2019. arXiv: 1907.00031. URL: https://arxiv.org/abs/1907. 00031.

Abstract: We introduce the thermodynamic variational objective (TVO) for learning in both continuous and discrete deep generative models. The TVO arises from a key connection between variational inference and thermodynamic integration that results in a tighter lower bound to the log marginal likelihood than the standard variational variational evidence lower bound (ELBO) while remaining as broadly applicable. We provide a computationally efficient gradient estimator for the TVO that applies to continuous, discrete, and non-reparameterizable distributions and show that the objective functions used in variational inference, variational autoencoders, wake sleep, and inference compilation are all special cases of the TVO. We use the TVO to learn both discrete and continuous deep generative models and empirically demonstrate state of the art model and inference network learning.

Le et al.: Revisiting Reweighted Wake-Sleep for Models with Stochastic Control Flow
T. A. Le, A. R. Kosiorek, N. Siddharth, Y. W. Teh, and F. Wood. “Revisiting Reweighted Wake-Sleep for Models with Stochastic Control Flow”. In: Proceedings of the conference on Uncertainty in Artificial Intelligence (UAI). 2019. arXiv: 1805.10469. URL: https://arxiv.org/abs/1805.10469.

Abstract: Stochastic control-flow models (SCFMs) are a class of generative models that involve branching on choices from discrete random variables. Amortized gradient-based learning of SCFMs is challenging as most approaches targeting discrete variables rely on their continuous relaxations—which can be intractable in SCFMs, as branching on relaxations requires evaluating all (exponentially many) branching paths. Tractable alternatives mainly combine REINFORCE with complex control-variate schemes to improve the variance of naive estimators. Here, we revisit the reweighted wake-sleep (RWS) (Bornschein and Bengio, 2015) algorithm, and through extensive evaluations, show that it outperforms current state-of-the-art methods in learning SCFMs. Further, in contrast to the importance weighted autoencoder, we observe that RWS learns better models and inference networks with increasing numbers of particles. Our results suggest that RWS is a competitive, often preferable, alternative for learning SCFMs.

We also performed research in the area of AutoML theory:

Bechavod et al.: Gaming helps! learning from strategic interactions in natural dynamics
Y. Bechavod, K. Ligett, S. Wu, and J. Ziani. “Gaming helps! learning from strategic interactions in natural dynamics”. In: International Conference on Artificial Intelligence and Statistics. PMLR. 2021, pp. 1234–1242.

Abstract: We consider an online regression setting in which individuals adapt to the regression model: arriving individuals are aware of the current model, and invest strategically in modifying their own features so as to improve the predicted score that the current model assigns to them. Such feature manipulation has been observed in various scenarios – from credit assessment to school admissions – posing a challenge for the learner. Surprisingly, we find that such strategic manipulations may in fact help the learner recover the meaningful variables – that is, the features that, when changed, affect the true label (as opposed to non-meaningful features that have no effect). We show that even simple behavior on the learner’s part allows her to simultaneously i) accurately recover the meaningful features, and ii) incentivize agents to invest in these meaningful features, providing incentives for improvement.

Hartford et al.: Valid causal inference with (some) invalid instruments
J. S. Hartford, V. Veitch, D. Sridhar, and K. Leyton-Brown. “Valid causal inference with (some) invalid instruments”. In: International Conference on Machine Learning. PMLR. 2021, pp. 4096–4106.

Abstract: Instrumental variable methods provide a powerful approach to estimating causal effects in the presence of unobserved confounding. But a key challenge when applying them is the reliance on untestable "exclusion" assumptions that rule out any relationship between the instrument variable and the response that is not mediated by the treatment. In this paper, we show how to perform consistent IV estimation despite violations of the exclusion assumption. In particular, we show that when one has multiple candidate instruments, only a majority of these candidates—or, more generally, the modal candidate-response relationship—needs to be valid to estimate the causal effect. Our approach uses an estimate of the modal prediction from an ensemble of instrumental variable estimators. The technique is simple to apply and is "black-box" in the sense that it may be used with any instrumental variable estimator as long as the treatment effect is identified for each valid instrument independently. As such, it is compatible with recent machine-learning based estimators that allow for the estimation of conditional average treatment effects (CATE) on complex, high dimensional data. Experimentally, we achieve accurate estimates of conditional average treatment effects using an ensemble of deep network-based estimators, including on a challenging simulated Mendelian Randomization problem.

Garg et al.: Learn to expect the unexpected: Probably approximately correct domain generalization
V. Garg, A. T. Kalai, K. Ligett, and S. Wu. “Learn to expect the unexpected: Probably approximately correct domain generalization”. In: International Conference on Artificial Intelligence and Statistics. PMLR. 2021, pp. 3574–3582.

Abstract: Domain generalization is the problem of machine learning when the training data and the test data come from different data domains. We present a simple theoretical model of learning to generalize across domains in which there is a meta-distribution over data distributions, and those data distributions may even have different supports. In our model, the training data given to a learning algorithm consists of multiple datasets each from a single domain drawn in turn from the meta-distribution. We study this model in three different problem settings—a multi-domain Massart noise setting, a decision tree multi-dataset setting, and a feature selection setting, and find that computationally efficient, polynomial-sample domain generalization is possible in each. Experiments demonstrate that our feature selection algorithm indeed ignores spurious correlations and improves generalization.

Kaplan et al.: Privately learning thresholds: Closing the exponential gap
H. Kaplan, K. Ligett, Y. Mansour, M. Naor, and U. Stemmer. “Privately learning thresholds: Closing the exponential gap”. In: Conference on Learning Theory. PMLR. 2020, pp. 2263–2285.

Abstract: We study the sample complexity of learning threshold functions under the constraint of differential privacy. It is assumed that each labeled example in the training data is the information of one individual and we would like to come up with a generalizing hypothesis h while guaranteeing differential privacy for the individuals. Intuitively, this means that any single labeled example in the training data should not have a significant effect on the choice of the hypothesis. This problem has received much attention recently; unlike the non-private case, where the sample complexity is independent of the domain size and just depends on the desired accuracy and confidence, for private learning the sample complexity must depend on the domain size X (even for approximate differential privacy). Alon et al. (STOC 2019) showed a lower bound of $\Omega(\log|X|)$ on the sample complexity and Bun et al. (FOCS 2015) presented an approximate-private learner with sample complexity $O\Lambda(2\log|X|)$. In this work we reduce this gap significantly,almost settling the sample complexity. We first present a new upper bound (algorithm) of $O\Lambda((\log|X|)2)$ on the sample complexity and then present an improved version with sample complexity $O((\log X)1.5)$. Our algorithm is constructed for the related interior point problem, where the goal is to find a point between the largest and smallest input elements.

It is based on selecting an input-dependent hash function and using it to embed the database into a domain whose size is reduced logarithmically; this results in a new database, an interior point of which can be used to generate an interior point in the original database in a differentially private manner.

Cameron et al.: Predicting propositional satisfiability via end-to-end learning

C. Cameron, R. Chen, J. Hartford, and K. Leyton-Brown. “Predicting propositional satisfiability via end-to-end learning”. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 34. 04. 2020, pp. 3324–3331.

Abstract: Strangely enough, it is possible to use machine learning models to predict the satisfiability status of hard SAT problems with accuracy considerably higher than random guessing. Existing methods have relied on extensive, manual feature engineering and computationally complex features (e.g., based on linear programming relaxations). We show for the first time that even better performance can be achieved by end-to-end learning methods — i.e., models that map directly from raw problem inputs to predictions and take only linear time to evaluate. Our work leverages deep network models which capture a key invariance exhibited by SAT problems: satisfiability status is unaffected by reordering variables and clauses. We showed that end-to-end learning with deep networks can outperform previous work on random 3-SAT problems at the solubility phase transition, where: (1) exactly 50% of problems are satisfiable; and (2) empirical runtimes of known solution methods scale exponentially with problem size (e.g., we achieved 84% prediction accuracy on 600-variable problems, which take hours to solve with state-of-the-art methods). We also showed that deep networks can generalize across problem sizes (e.g., a network trained only on 100-variable problems, which typically take about 10 ms to solve, achieved 81% accuracy on 600-variable problems).

Weisz et al.: ImpatientCapsAndRuns: Approximately Optimal Algorithm Configuration from an Infinite Pool

G. Weisz, A. György, W.-I. Lin, D. R. Graham, K. Leyton-Brown, C. Szepesvari, and B. Lucier. “ImpatientCapsAn- dRuns: Approximately Optimal Algorithm Configuration from an Infinite Pool”. In: NeurIPS. 2020.

Abstract: Algorithm configuration procedures optimize parameters of a given algorithm to perform well over a distribution of inputs. Recent theoretical work focused on the case of selecting between a small number of alternatives. In practice, parameter spaces are often very large or infinite, and so successful heuristic procedures discard parameters impatiently”, based on very few observations. Inspired by this idea, we introduce ImpatientCapsAndRuns, which quickly discards less promising configurations, significantly speeding up the search procedure compared to previous algorithms with theoretical guarantees, while still achieving optimal runtime up to logarithmic factors under mild assumptions. Experimental results demonstrate a practical improvement.

Shenfeld et al.: A necessary and sufficient stability notion for adaptive generalization
M. Shenfeld and K. Ligett. “A necessary and sufficient stability notion for adaptive generalization”. In: Advances in Neural Information Processing Systems 32 (2019), pp. 11485–11494.

Abstract: We introduce a new notion of the stability of computations, which holds under post-processing and adaptive composition. We show that the notion is both necessary and sufficient to ensure generalization in the face of adaptivity, for any computations that respond to bounded-sensitivity linear queries while providing accuracy with respect to the data sample set. The stability notion is based on quantifying the effect of observing a computation’s outputs on the posterior over the data sample elements. We show a separation between this stability notion and previously studied notion and observe that all differentially private algorithms also satisfy this notion.

Jung et al.: A New Analysis of Differential Privacy’s Generalization Guarantees
C. Jung, K. Ligett, S. Neel, A. Roth, S. Sharifi-Malvajerdi, and M. Shenfeld. “A New Analysis of Differential Privacy’s Generalization Guarantees”. In: arXiv preprint arXiv:1909.03577 (2019).

Abstract: We give a new proof of the "transfer theorem" underlying adaptive data analysis: that any mechanism for answering adaptively chosen statistical queries that is differentially private and sample-accurate is also accurate out-of- sample. Our new proof is elementary and gives structural insights that we expect will be useful elsewhere. We show: 1) that differential privacy ensures that the expectation of any query on the posterior distribution on datasets induced by the transcript of the interaction is close to its true value on the data distribution, and 2) sample accuracy on its own ensures that any query answer produced by the mechanism is close to its posterior expectation with high probability. This second claim follows from a thought experiment in which we imagine that the dataset is resampled from the posterior distribution after the mechanism has committed to its answers. The transfer theorem then follows by summing these two bounds, and in particular, avoids the "monitor argument" used to derive high probability bounds in prior work. An upshot of our new proof technique is that the concrete bounds we obtain are substantially better than the best previously known bounds, even though the improvements are in the constants, rather than the asymptotics (which are known to be tight). As we show, our new bounds outperform the naive "sample-splitting" baseline at dramatically smaller dataset sizes compared to the previous state of the art, bringing techniques from this literature closer to practicality.

Cai et al.: Third-party data providers ruin simple mechanisms
Y. Cai, F. Echenique, H. Fu, K. Ligett, A. Wierman, and J. Ziani. “Third-party data providers ruin simple mechanisms”. In: Proceedings of the ACM on Measurement and Analysis of Computing Systems 4.1 (2020), pp. 1–31.

Abstract: Motivated by the growing prominence of third-party data providers in online marketplaces, this paper studies the impact of the presence of third-party data providers on mechanism design. When no data provider is present, it has been shown that simple mechanisms are "good enough" – they can achieve a constant fraction of the revenue of optimal mechanisms. The results in this paper demonstrate that this is no longer true in the presence of a third-party data provider who can provide the bidder with a signal that is correlated with the item type. Specifically, even with a single seller, a single bidder, and a single item of uncertain type for sale, the strategies of pricing each item-type separately (the analog of item pricing for multi-item auctions) and bundling all item-types under a single price (the analog of grand bundling) can both simultaneously be a logarithmic factor worse than the optimal revenue. Further, in the presence of a data provider, item-type partitioning mechanisms—a more general class of mechanisms which divide item-types into disjoint groups and offer prices for each group—still cannot achieve within a loglog factor of the optimal revenue. Thus, our results highlight that the presence of a data-provider forces the use of more complicated mechanisms in order to achieve a constant fraction of the optimal revenue.

Ligett et al.: Bounded-leakage differential privacy
K. Ligett, C. Peale, and O. Reingold. "Bounded-leakage differential privacy". In: 1st Symposium on Foundations of Responsible Computing (FORC 2020). Schloss Dagstuhl-Leibniz-Zentrum für Informatik. 2020.

Abstract: We introduce and study a relaxation of differential privacy [Dwork et al., 2006] that accounts for mechanisms that leak some additional, bounded information about the database. We apply this notion to reason about two distinct settings where the notion of differential privacy is of limited use. First, we consider cases, such as in the 2020 US Census [Abowd, 2018], in which some information about the database is released exactly or with small noise. Second, we consider the accumulation of privacy harms for an individual across studies that may not even include the data of this individual. The tools that we develop for bounded-leakage differential privacy allow us reason about privacy loss in these settings, and to show that individuals preserve some meaningful protections.

Finally, we developed two systems during the course of this program, Daphne and PyProb. Our work in this area started with LF-PPL:

Zhou et al.: LF-PPL: A Low-Level First Order Probabilistic Programming Language for Non-Differentiable Models
Y. Zhou, B. J. Gram-Hansen, T. Kohn, T. Rainforth, H. Yang, and F. Wood. "LF-PPL: A Low-Level First Order Probabilistic Programming Language for Non-Differentiable Models". In: Proceedings of the Twentieth International Conference on Artificial Intelligence and Statistics (AISTATS). 2019. arXiv: 1903.02482. URL: https://arxiv.org/pdf/1903.02482.pdf.

Abstract: We develop a new Low-level, First-order Probabilistic Programming Language (LF-PPL) suited for models containing a mix of continuous, discrete, and/or piecewise-continuous variables. The key success of this language and its compilation scheme is in its ability to automatically distinguish parameters the density function is discontinuous with respect to, while further providing runtime checks for boundary crossings. This enables the introduction of new inference engines that are able to exploit gradient information, while remaining efficient for models which are not everywhere differentiable. We demonstrate this ability by incorporating a discontinuous Hamiltonian Monte Carlo (DHMC) inference engine that is able to deliver automated and efficient inference for non-differentiable models. Our system is backed up by a mathematical formalism that ensures that any model expressed in this language has a density with measure zero discontinuities to maintain the validity of the inference engine.

Hasty v2 = Daphne,[7] is the subject of "Structured Conditional Continuous Normalizing Flows for Efficient Amortized Inference in Graphical Models" [9]. From an AutoML perspective, Daphne constructs inverse graphical model structures from graphical model generative models specified as programs. Further, in graphical models with only continuous random variables, Daphne automatically provisions an architecture for an amortized inference artifact in the form of a flow and trains it, again automatically, to perform amortized inference. To use Daphne to produce AutoML primitives one must program a domain-specific generative model then wait for an artifact that performs conditional inference in this model automatically. This amortized inference artifact is the model-based AutoML featurization primitive. The language Daphne compiles is a bespoke graphical model domain specific language. The results of our work are summarized in:

Weilbach et al.: Structured Conditional Continuous Normalizing Flows for Efficient Amortized Inference in Graphical Models
C. Weilbach, B. Beronov, F. Wood, and W. Harvey. "Structured Conditional Continuous Normalizing Flows for Efficient Amortized Inference in Graphical Models". In: Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics (AISTATS). PMLR 108:4441-4451. 2020, pp. 4441–4451. URL: http://proceedings. mlr.press/v108/weilbach20a.html.

Abstract: We exploit minimally faithful inversion of graphical model structures to specify sparse continuous normalizing flows (CNFs) for amortized inference. We find that the sparsity of this factorization can be exploited to reduce the numbers of parameters in the neural network, adaptive integration steps of the flow, and consequently FLOPs at both training and inference time without decreasing performance in comparison to unconstrained flows. By expressing the structure inversion as a compilation pass in a probabilistic programming language, we are able to apply it in a novel way to models as complex as convolutional neural networks. Furthermore, we extend the training objective for CNFs in the context of inference amortization to the symmetric Kullback-Leibler divergence, and demonstrate its theoretical and practical advantages.

Hasty v4 = PyProb,[8] a simulator agnostic tool developed for inference in simulators has been applied to high energy physics [3] (best paper finalist), non-invasive assessment of composite material properties during manufacture [10, 11], and to understanding COVID-19 dynamics [12]. From an AutoML perspective, PyProb constructs a generic neural network that guides amortized inference in generative models specified as programs. To use PyProb to produce AutoML primitives one must program a domain-specific generative model then wait for PyProb to train an artifact that performs conditional inference in this model automatically. This amortized inference artifact is the model-based AutoML featurization primitive. PyProb differs from Daphne in that a "normal" programming language can be used to write the generative model rather than a restrictive graphical model DSL. The results of our work are summarized in:

Munk et al.: Deep probabilistic surrogate networks for universal simulator approximation
A. Munk, A. S´cibior, A. Baydin, A. Stewart, A. Fernlund, A. Poursartip, and F. Wood. "Deep probabilistic surrogate net- works for universal simulator approximation". In: The second International Conference on Probabilistic Programming (PROBPROG). 2020. arXiv:

[7] `https://github.com/plai-group/daphne`

[8] https://github.com/plai-group/pyprob

1910.11950. URL: https://arxiv.org/abs/1910.11950.

Abstract: We present a framework for automatically structuring and training fast, approximate, deep neural surrogates of existing stochastic simulators. Unlike traditional approaches to surrogate modeling, our surrogates retain the interpretable structure of the reference simulators. The particular way we achieve this allows us to replace the reference simulator with the surrogate when undertaking amortized inference in the probabilistic programming sense. The fidelity and speed of our surrogates allow for not only faster "forward" stochastic simulation but also for accurate and substantially faster inference. We support these claims via experiments that involve a commercial composite-materials curing simulator. Employing our surrogate modeling technique makes inference an order of magnitude faster, opening up the possibility of doing simulator-based, non-invasive, just-in-time parts quality testing; in this case inferring safety-critical latent internal temperature profiles of composite materials undergoing curing from surface temperature profile measurements.

Naderiparizi et al.: Amortized rejection sampling in universal probabilistic programming

Naderiparizi, A. S´cibior, A. Munk, M. Ghadiri, A. Baydin, B. Gram-Hansen, C. Schroeder de Witt, R. Zinkov, P. Torr, Rainforth, Y. Whye Teh, and F. Wood. “Amortized rejection sampling in universal probabilistic programming”. 2019. URL: https://arxiv.org/abs/1910.09056.

Abstract: Existing approaches to amortized inference in probabilistic programs with unbounded loops can produce estimators with infinite variance. An instance of this is importance sampling inference in programs that explicitly include rejection sampling as part of the user-programmed generative procedure. In this paper we develop a new and efficient amortized importance sampling estimator. We prove finite variance of our estimator and empirically demonstrate our method’s correctness and efficiency compared to existing alternatives on generative programs containing rejection sampling loops and discuss how to implement our method in a generic probabilistic programming framework.

Baydin et al.: Etalumis: Bringing Probabilistic Programming to Scientific Simulators at Scale

A. G. Baydin, L. Shao, W. Bhimji, L. Heinrich, L. Meadows, J. Liu, A. Munk, S. Naderiparizi, B. Gram-Hansen, G. Louppe, et al. “Etalumis: Bringing Probabilistic Programming to Scientific Simulators at Scale”. In: the International Conference for High Performance Computing, Networking, Storage and Analysis (SC ’19). 2019. arXiv: 1907.03382. URL: https://dl.acm.org/doi/10.1145/3295500.3356180.

Abstract: Probabilistic programming languages (PPLs) are receiving widespread attention for performing Bayesian inference in complex generative models. However, applications to science remain limited because of the impracticability of rewriting complex scientific simulators in a PPL, the computational cost of inference, and the lack of scalable implementations. To address these, we present a novel PPL framework that couples directly to existing scientific simulators through a cross-platform probabilistic execution protocol and provides Markov chain Monte Carlo (MCMC) and deep-learning-based inference compilation (IC) engines for tractable inference. To guide IC inference, we perform distributed training of a dynamic 3DCNN–LSTM architecture with a PyTorch-MPI-based framework on 1,024 32-core CPU nodes of the Cori supercomputer with a global minibatch size of 128k: achieving a performance of 450 Tflop/s through enhancements to PyTorch. We demonstrate a Large Hadron Collider (LHC) use-case with the C++ Sherpa simulator and achieve the largest-scale posterior inference in a Turing-complete PPL.

Baydin et al.: Efficient Probabilistic Inference in the Quest for Physics Beyond the Standard Model

A. G. Baydin, L. Heinrich, W. Bhimji, B. Gram-Hansen, G. Louppe, L. Shao, K. Cranmer, F. Wood, et al. "Efficient Probabilistic Inference in the Quest for Physics Beyond the Standard Model". In: Thirty-second Conference on Neural Information Processing Systems (NeurIPS). 2019. arXiv: 1807.07706. URL: https://papers.nips.cc/paper/ 8785 - efficient - probabilistic - inference - in - the - quest - for - physics - beyond - the - standard - model.

Abstract: We present a novel probabilistic programming framework that couples directly to existing large-scale simulators through a cross-platform probabilistic execution protocol, which allows general-purpose inference engines to record and control random number draws within simulators in a language-agnostic way. The execution of existing simulators as probabilistic programs enables highly interpretable posterior inference in the structured model defined by the simulator code base. We demonstrate the technique in particle physics, on a scientifically accurate simulation of the tau lepton decay, which is a key ingredient in establishing the properties of the Higgs boson. Inference efficiency is achieved via inference compilation where a deep recurrent neural network is trained to parameterize proposal distributions and control the stochastic simulator in a sequential importance sampling scheme, at a fraction of the computational cost of a Markov chain Monte Carlo baseline.

4.2.4 Other Work

Other research generated by sponsored HQP but not directly related to AutoML includes:

Warrington et al.: Robust Asymmetric Learning in POMDPs

A. Warrington, J. W. Lavington, A. Scibior, M. Schmidt, and F. Wood. "Robust Asymmetric Learning in POMDPs". 2020. URL: https://arxiv.org/abs/2012.15566.

Abstract: Policies for partially observed Markov decision processes can be efficiently learned by imitating policies for the corresponding fully observed Markov decision processes. Unfortunately, existing approaches for this kind of imitation learning have a serious flaw: the expert does not know what the trainee cannot see, and so may encourage actions that are sub-optimal, even unsafe, under partial information. We derive an objective to instead train the expert to maximize the expected reward of the imitating agent policy, and use it to construct an efficient algorithm, adaptive asymmetric DAgger (A2D), that jointly trains the expert and the agent. We show that A2D produces an expert policy that the agent can safely imitate, in turn outperforming policies learned by imitating a fixed expert.

Harvey et al.: Near-Optimal Glimpse Sequences for Improved Hard Attention Neural Network Training

W. Harvey, M. Teng, and F. Wood. "Near-Optimal Glimpse Sequences for Improved Hard Attention Neural Network Training". 2019. URL: https://arxiv.org/abs/1906.05462.

Abstract: Hard visual attention is a promising approach to reduce the computational burden of modern computer vision methodologies. Hard attention mechanisms are typically non-differentiable. They can be trained with reinforcement learning but the high-variance training this entails hinders more widespread application. We show how hard attention for image classification can be framed as a Bayesian optimal experimental design (BOED) problem. From this perspective, the optimal locations to attend to are those which provide the greatest expected reduction in the entropy of the classification distribution. We introduce methodology from the BOED literature to approximate this optimal behaviour, and use it to generate ‘near-optimal’ sequences of attention locations. We then show how to use such sequences to partially supervise, and therefore speed up, the training of a hard attention mechanism. Although generating these sequences is computationally expensive, they can be reused by any other networks later trained on the same task.

Wood et al.: Planning as Inference in Epidemiological Models

F. Wood, A. Warrington, S. Naderiparizi, C. Weilbach, V. Masrani, W. Harvey, A. Scibior, B. Beronov, and A. Nasseri. “Planning as Inference in Epidemiological Models”. 2020. URL: https://arxiv.org/abs/2003.13221.

Abstract: In this work we demonstrate how existing software tools can be used to automate parts of infectious disease- control policy-making via performing inference in existing epidemiological dynamics models. The kind of inference tasks undertaken include computing, for planning purposes, the posterior distribution over putatively controllable, via direct policy-making choices, simulation model parameters that give rise to acceptable disease progression outcomes. Neither the full capabilities of such inference automation software tools nor their utility for planning is widely disseminated at the current time. Timely gains in understanding about these tools and how they can be used may lead to more fine-grained and less economically damaging policy prescriptions, particularly during the current COVID-19 pandemic.

Warrington et al.: The Virtual Patch Clamp: Imputing C. elegans Membrane Potentials from Calcium Imaging

A. Warrington, A. Spencer, and F. Wood. “The Virtual Patch Clamp: Imputing C. elegans Membrane Potentials from Calcium Imaging”. In: NeurIPS 2019 Workshop Neuro AI. 2019. arXiv: 1907.11075. URL: https://arxiv.org/ pdf/1907.11075.pdf.

Abstract: We develop a stochastic whole-brain and body simulator of the nematode roundworm Caenorhabditis elegans (C. elegans) and show that it is sufficiently regularizing to allow imputation of latent membrane potentials from partial calcium fluorescence imaging observations. This is the first attempt we know of to "complete the circle," where an anatomically grounded whole-connectome simulator is used to impute a time-varying "brain" state at single-cell fidelity from covariates that are measurable in practice. The sequential Monte Carlo (SMC) method we employ not only enables imputation of said latent states but also presents a strategy for learning simulator parameters via variational optimization of the noisy model evidence approximation provided by SMC. Our imputation and parameter estimation experiments were conducted on distributed systems using novel implementations of the aforementioned techniques applied to synthetic data of dimension and type representative of that which are measured in laboratories currently.

4.3 Software and Datasets Contributed to D3M Codebase

UBC contributed 18 primitives (with sample pipelines) and 1 Dataset to the D3M program, which are accessible on the D3M Gitlab and on UBC PLAI Group Github.[9] These primitives include:

1. Canonical Correlation Forests (CCF) Classifier and Regressor: CCF [13] is a new decision tree ensemble method which naturally accommodates multiple outputs, provides a similar computational complexity to random forests, and inherits their impressive robustness to the choice of input parameters. It uses semantic types to determine which columns to operate on. We contributed a classifier and a regressor based on the CCF method to the D3M program.
2. Principal Component Analysis: The PCA primitive is used to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. It uses Singular Value Decomposition of the data to project it to a lower dimensional space.
3. K-Means: The K-Means algorithm clusters data by trying to separate samples in n groups of equal variance, minimizing a criterion known as the inertia or within-cluster sum-of-squares. This algorithm requires the number of clusters to be specified. It scales well to large number of samples and has been used across a large range of application areas in many different fields.
4. Bayesian Linear-Regression: A Bayesian treatment of linear regression.
5. Diagonal Multivariate Normal: This primitive basedallows fitting, and sampling from a multivariate Gaussian (with diagonal covariance matrix).
6. Semantic Type Inference: This primitive can detect the semantic type of inputed column data. It currently supports 78 Semantic Types [14].
7. Deep Markov Model: A deep markov model in the d3m interface implemented in PyTorch.
8. Multilayer Perceptron Classifier and Regressor: We contributed a feed-forward neural network classification and regressor primitive developed using PyTorch to the program. It can be configured with input and output dimensions, number of layers (depth), and number of units in each layer except the last one (width).
9. Convolutional Neural Network (CNN): We contributed an implementation of CNNs using the PyTorch framework, used to extract deep features from images. It can be used as a pre-trained feature extractor, to extract features from convolutional layers or the fully connected layers, or fine-tuned to fit (classification/regression) new data. Available pre-trained CNN models include: VGG-16, VGG-16 with Batch-Norm, GoogLeNet, ResNet-34, and MobileNet, all pre-trained on ImageNet.
10. In addition, we included seperate implementations of GoogleNet [15], MobileNet [16], ResNet [17], and VGGNet [18] implemented in PyTorch which allow for faster installation and search. These separate primitves are re-trained on ImageNet, and can be used as a pre-trained feature extractor, to extract features from convolutional layers or be fine-tunned to fit new data.
11. Simple-CNAPS: Simple CNAPS [6, 7] is a simple classcovariance-based distance metric, namely the Maha- lanobis distance, adopted into a state-of-the-art few-shot learning approach called CNAPS, which leads to a significant performance improvement. It is able to learn adaptive feature extractors that allow useful estimation of the high dimensional feature

[9] https://github.com/plai-group/ubc_primitives

covariances required by this metric from few samples.

12. Phone Parser: A primitive that can parse phone numbers using inference compilation, and is used as a demo of our PyProb-based primitives.

We also contributed a metadataset dataset accompanying the SimpleCNAPS primitive alongside the Quick-Start Guide (in collaboration with Arrayfire) and the advanced tutorial in the D3M Documentation.

5.0 Conclusions

While a tremendous amount of research, innovation, and training was accomplished by this group under this funding and through participation in this program, the overall objective of the program fell short of its potential. We believe this is because of several major things. One, industrial interest in AutoML grew significantly in parallel to the evolution of the program. This meant that there was stiff AutoML competition which is better organized and frankly better resourced in terms of compute, a key ingredient for success in this space. Two, the program didn't run long enough to fully integrate all of the ideas from its performers. In particular, the custom featurization primitives our team worked towards contributing were not ultimately embedded in any of the AutoML systems produced within this program. This resulted in a major handicapping of the systems both in terms of the problem types they could address and probably also in terms of the quality of results on customer problems. And three, the evaluation techniques used in this program concentrated effort almost solely on the pipeline search problem, with problem types and performance evaluation being dictated by a concrete set of problems frankly well-served by existing open-source and industrial AutoML systems. This left very little appetite or incentive for performers to truly expand AutoML capabilities; the pipeline search problem being nearly too challenging by itself anyway. AutoML as a fascinating problem will not go away anytime soon and we encourage the formation of a new program whose goals are not only as audacious as the original D3M goals, but remain so throughout the evolution of the program.

6.0 References

[1] Christian Weilbach, Boyan Beronov, William Harvey, and Frank Wood. "Structured Conditional Continuous Normalizing Flows for Efficient Amortized Inference in Graphical Models". In: *The 23rd International Confer- ence on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]*. Ed. by Silvia Chiappa and Roberto Calandra. Vol. 108. Proceedings of Machine Learning Research. PMLR, 2020, pp. 4441–4451. URL: http://proceedings.mlr.press/v108/weilbach20a.html.

[2] Tuan Anh Le, Atilim Gunes Baydin, and Frank Wood. "Inference compilation and universal probabilistic programming". In: *Artificial Intelligence and Statistics*. PMLR. 2017, pp. 1338–1348.

[3] Atılım Günes¸ Baydin, Lei Shao, Wahid Bhimji, Lukas Heinrich, Lawrence Meadows, Jialin Liu, Andreas Munk, Saeid Naderiparizi, Bradley Gram-Hansen, Gilles Louppe, et al. "Etalumis: Bringing Probabilistic Programming to Scientific Simulators at Scale". In: *the International Conference for High Performance Computing, Networking, Storage and Analysis (SC '19)*. 2019. arXiv: 1907.03382. URL: https://dl.acm.org/doi/10.1145/ 3295500.3356180.

[4] Atilim Gunes Baydin, Lukas Heinrich, Wahid Bhimji, Bradley Gram-Hansen, Gilles Louppe, Lei Shao, Kyle Cranmer, Frank Wood, et al. "Efficient Probabilistic Inference in the Quest for Physics Beyond the Standard Model". In: *Thirty-second Conference on Neural Information Processing Systems (NeurIPS)*. 2019. arXiv: 1807.07706. URL: https://papers.nips.cc/paper/8785-efficient-probabilistic-inference-in-the-quest-for-physics-beyond-the-standard-model.

[5] Jason Yoo, Tony Joseph, Dylan Yung, S. Ali Nasseri, and Frank Wood. "Ensemble Squared: A Meta AutoML System". 2020. URL: https://arxiv.org/abs/2012.05390.

[6] Peyman Bateni, Raghav Goyal, Vaden Masrani, Frank Wood, and Leonid Sigal. "Improved Few-Shot Visual Classification". In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020. arXiv: 1912.03432. URL: https://openaccess.thecvf.com/content_CVPR_2020/html/Bateni_Improved_Few-Shot_Visual_Classification_CVPR_2020_paper.html.

[7] Peyman Bateni, Jarred Barber, Jan-Willem van de Meent, and Frank Wood. "Improving Few-Shot Visual Classification with Unlabelled Examples". 2020. URL: https://arxiv.org/abs/2006.12245.

[8] P. Gijsbers, E. LeDell, S. Poirier, J. Thomas, B. Bischl, and J. Vanschoren. "An Open Source AutoML Bench- mark". In: *arXiv preprint arXiv:1907.00909 [cs.LG]* (2019). Accepted at AutoML Workshop at ICML 2019. URL: https://arxiv.org/abs/1907.00909.

[9] Christian Weilbach, Boyan Beronov, Frank Wood, and William Harvey. "Structured Conditional Continuous Normalizing Flows for Efficient Amortized Inference in Graphical Models". In: *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics (AISTATS)*. PMLR 108:4441-4451. 2020, pp. 4441–4451. URL: http://proceedings.mlr.press/v108/weilbach20a.html.

[10] Andreas Munk, Adam Ścibior, AG Baydin, A Stewart, A Fernlund, A Poursartip, and Frank Wood. "Deep probabilistic surrogate networks for universal simulator approximation". In: *The second International Conference on Probabilistic Programming (PROBPROG)*. 2020. arXiv:

1910.11950. URL: https://arxiv.org/abs/ 1910.11950.

[11] Saeid Naderiparizi, Adam Ścibior, Andreas Munk, Mehrdad Ghadiri, Atılım Güneş Baydin, Bradley Gram- Hansen, Christian Schroeder de Witt, Robert Zinkov, Philip HS Torr, Tom Rainforth, et al. "Amortized rejection sampling in universal probabilistic programming". In: *International Conference on Probabilistic Programming (PROBPROG)*. 2020. arXiv: 1910.09056. URL: https://arxiv.org/abs/1910.09056.

[12] Frank Wood, Andrew Warrington, Saeid Naderiparizi, Christian Weilbach, Vaden Masrani, William Harvey, Adam Scibior, Boyan Beronov, and Ali Nasseri. "Planning as Inference in Epidemiological Models". 2020. URL: https://arxiv.org/abs/2003.13221.

[13] Tom Rainforth and Frank Wood. "Canonical correlation forests". In: *arXiv preprint arXiv:1507.05444* (2015).

[14] Madelon Hulsebos, Kevin Hu, Michiel Bakker, Emanuel Zgraggen, Arvind Satyanarayan, Tim Kraska, Çagatay Demiralp, and César Hidalgo. "Sherlock: A deep learning approach to semantic data type detection". In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2019, pp. 1500–1508.

[15] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. "Going deeper with convolutions". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 1–9.

[16] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco An- dreetto, and Hartwig Adam. "Mobilenets: Efficient convolutional neural networks for mobile vision applications". In: *arXiv preprint arXiv:1704.04861* (2017).

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[18] Karen Simonyan and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556* (2014).