Michael C. Hughes Research Statement www.michaelchughes.com

I develop machine learning methods that discover actionable knowledge from large, messy datasets. My applied focus is healthcare, where novel probabilistic models and optimization algorithms are needed to learn from big, noisy, unlabeled datasets to produce recommendations clinicians can trust. In one project, we modeled vital sign trajectories in the Intensive Care Unit across 36,000 patients to predict the need for mechanical ventilators. In ongoing work with psychiatrists at Massachusetts General Hospital, we are finding subpopulations of major depression disorder which require distinct drug combination therapies. My research provides better methods to the machine learning community, patient-specific treatment advice to doctors, and scientific insight into possible subtypes or trajectories of targeted diseases. I strive to deliver results beyond top-tier publications, including open-source software and (eventually) deployed clinical decision support systems.

My research agenda addresses three fundamental challenges in machine learning (ML) in order to deliver the promise of probabilistic models to the broader scientific and medical community:

Q1: How can we *combine abundant unlabeled data with rare task-specific labels*? Observational medical datasets contain many thousands of patients, but only a few will have reliable labels of treatments that were *successful* (not just attempted). Because acquiring labels is expensive, we must develop a *semi-supervised* approach that learns from few labeled examples *and* many unlabeled ones. Latent variable models which explain both the data x and labels y can achieve two needed goals: good generative models of data x (clinical insight on possible subtypes) and good prediction of labels y from data x (personalized treatments that work). However, despite decades of work we find that existing methods always fail on at least one of these goals. We have developed a new *prediction-constrained* training objective, which enables the practioner to balance both goals. Clinicians can thus train and deploy models that offer the best possible insights into patient clusters or disease subtypes while also satisfying a predefined accuracy guarantee for treatment prediction.

Q2: How can we *improve inference algorithms* to reach good solutions consistently? Models that are flexible enough for big clinical data, such as Bayesian nonparametric (BNP) models or deep neural networks, suffer from unreliable training. Typical methods optimize non-convex objective functions via iterative updates to subsets of parameters. Even on small data, random initializations become stuck at poor local optima and fail to reach even the best mode, let alone explore the real posterior. Scaling to millions of examples only exaggerates this problem. My Ph.D. work on reliable BNP suggests that standard updates can be interleaved with *data-driven proposals* that create new clusters to better explain some examples or remove irrelevant clusters. Our proposal-driven algorithms reach qualitatively better solutions without expensive restarts or cross-validation.

Q3: How can we *optimize models to be interpretable* to human users? Large, flexible models are needed for high accuracy predictions on complex medical datasets, but how can users comprehend and trust their predictions? While many efforts interpret fixed models after they are trained, our recent work looks at *optimizing* models to be more interpretable. We have trained some models to be *easier to simulate*, so doctors could quickly step through the prediction process to understand how a specific input leads to the provided output. We have also thought about counter-factual reasoning: understanding how slightly perturbed inputs lead to different predictions. Clinicians have stressed that answers to these questions are essential to trusting an ML system as part of a real patient's treatment.

Semi-Supervised Models for Personalized Medicine: Progress on Q1

Personalized healthcare is a natural problem for latent variable models, which can infer patient-specific low-dimensional representations useful for task-specific predictions. Clinicians often prefer models where hidden variables h (e.g. trajectory from severe kidney malfunction to healthy state) jointly explain input data x (e.g. labs, vital signs) and output labels y (e.g. prescribed interventions). Our paper at the 2017 Joint Summits of the American Medical Informatics Association (AMIA) applied auto-regressive switching-state models to time-series data from 36,000 ICU patients [4]. Although trained in an *unsupervised* way (discovering common trajectories without labels y), our learned representations improved 4-hour look-ahead predictions of need for mechanical ventilation. Latent variable models can improve personalized care and smooth logistics in the busy ICU, but to reach higher accuracies we must *supervise* training by including labels.

Despite decades of work on (semi-)supervision of latent variable models, existing methods remain unsatisfactory. For example, a supervised topic model survey [5] shows no accuracy gains over unsupervised models when predicting patient outcomes from clinical notes. My postdoc work has discovered why: previous training objectives do not prioritize predicting y from x alone. Instead, our recent preprint [14] shows many previous objectives [10; 22; 23; 3] can be formally reduced to optimizing a *joint* probability $p(x, y)$, where the labels y are replicated $\lambda \geq 1$ times. When the model is (inevitably) misspecified for real data, such training may not predict y from x well.

To improve accuracy, our proposed prediction-constrained (PC) objective directly delivers the best possible generative model that meets a provided quality guarantee on the model's y from x predictions. Our PC-trained mixture and topic models reach qualitatively better solutions on toy examples with misspecification and text sentiment analysis. Unlike purely discriminative approaches, PC training can improve predictions by including many unlabeled examples, while boosting data likelihoods as well. This work is under review at a major ML conference. Working with Dr. Perlis and Dr. McCoy at Massachusetts General Hospital, we have applied PC training to recommend antidepressants for patients with major depression, resulting in two papers in the NIPS Workshop on ML for Health [18; 17] and a clinical journal publication in preparation for JAMA Psychiatry.

Future work: extensions to time-series models and reinforcement learning. Our prediction-constrained (PC) framework should translate usefully to many currently unsupervised latent variable models, including our ICU time-series applications and recent compositions of linear dynamical systems with flexible neural net likelihoods [8]. Further ahead, prediction-constrained methods could help reinforcement learning models make sequential drug recommendations that maximize reward even when labels for action success are rare.

Future work: prediction-constrained posteriors to handle uncertainty. Our current PC framework delivers point estimates of parameters. We would rather estimate *posterior distributions* and thus manage our uncertainty. However, standard variational methods for estimating approximate posteriors are not tractable for PC training. These methods optimize a *lower bound* on the data evidence $\log p(x)$ but one term in our PC objective requires an *upper bound* instead. Such upper bounds could also be useful for other common questions, such as estimating $p(\text{validation data}|\text{train data})$ for a given model. Estimating variational upper bounds remains a challenging methodological problem. Ultimately, including uncertainty in our PC-trained models could provide better calibrated suggestions for interventions, especially when test examples diverge from training data.

Reliable Inference for Bayesian Nonparametrics: Progress on Q2

My Ph.D. thesis developed reliable optimization methods for Bayesian nonparametric (BNP) clustering models, particularly the Dirichlet process (DP) and its extensions to time series. While parametric models like k-means require an *a priori* number of clusters, BNP models *learn* the number of clusters needed to explain a dataset, balancing gains in quality from adding more clusters with a rich-get-richer preference for fewer clusters. BNP models thus promise an automatic solution to the model selection problem in one training run that avoids expensive cross-validation or fancy initializations. However, standard BNP algorithms do not fulfill this promise. Both Markov chain Monte Carlo (MCMC) methods and optimization-based variational inference use restrictive update steps that get stuck in poor local optima due to the limited range of each update.

Our 2013 Neural Information Processing Systems (NIPS) conference paper [12] developed a new algorithm for BNP mixture models that used data-driven proposals to jump out of local optima by adding crucial missing clusters or removing redundant clusters. These proposals optimize a variational objective function which tightly bounds the marginal likelihood and thus exhibits the "Ockham's razor" effect that penalizes models that are too complex or too simple. Furthermore, our method scales to large datasets by processing data one small batch at a time. Unlike stochastic methods that require tuning a nuisance learning rate [6], our scalable memoized algorithm has no learning rate at all yet guarantees that the objective will monotonically increase after every step.

Later, we extended this approach to topic modeling with the hierarchical Dirichlet process (HDP) [15], sequence segmentation via the HDP hidden Markov model (HDP-HMM) [16], and compositional models for natural images [7]. These settings are challenging due to non-conjugacy and tighter data dependencies. Nevertheless, we can optimize a variational lower bound via data-driven proposal moves that scale to millions of NY Times articles or the entire human genome. To make these contributions accessible, I released BNPy, an open-source Python software package [11] now used by many researchers, including data science teams at the New York Times.

Future work: Sparsity and recognition networks for extreme scalability. Two major challenges prevent BNP clustering from scaling to billions of examples and thousands of clusters. First, the bottleneck of training is the runtime cost of fitting a large model to each new example. Recent variational auto-encoders [9; 19] pursue an overall Bayesian variational objective but use a fast, feed-forward neural network to approximate the posterior needed for each new example. Thus, information from previous examples can help cluster new data faster and thus amortize costs. Second, our recent preprint [13] suggests that *sparse* representations of per-example posteriors can reduce storage and improve speed, making BNP models with thousands of clusters possible. Incorporating these ideas together with data-driven proposal moves and our semi-supervised PC training remains an open problem with huge potential for discovering thousands of prototypical disease subtypes or patient trajectories from million-patient, multi-hospital datasets.

Future work: Guarantees on proposal quality. Thus far, the data-driven proposals used in our algorithms have been heuristically designed and have no guarantees about how far its final estimate is from a global optima. Connections to the approximation quality guarantees in the *k-means++* algorithm [2] might be possible with BNP variational objectives, because all these objectives are instances of minimizing (expected) Bregman divergences between cluster centers and data points [1]. Proposal moves that deliver guaranteed approximation ratios to the best-possible clustering model would be a huge advancement for the reliability of BNP on large datasets.

Optimizing Deep Models for Interpretability: Progress on Q3

While deep learning has made impressive strides on image and language tasks, many clinical practitioners are reluctant to adopt deep models because their predictions are difficult to interpret, explain, and trust. During my postdoc, I've pursued two efforts to close this trust gap.

First, our IJCAI 2017 paper [20] shows how to train models to respect expert annotations of features relevant to specific examples. Furthermore, even without expert annotations our method can actively *discover* models with different per-example decision rationales via the same regularization process. This gives practitioners valuable tools to debug pretrained models or discover confounding features hiding in their datasets.

Second, in a paper recently accepted to the 2018 Association for the Advancement of Artificial Intelligence (AAAI) conference [21], we introduce *tree-regularization* as a method to optimize deep models so they are *human-simulatable*. Small decision trees with only a few nodes make it easy for humans to step through the entire prediction process, and thus enjoy widespread use in manual medical diagnosis. In contrast, feed-forward networks with a dozen hidden units can have far too many parameters and connections for a human to simulate. Deep models for sequences are even more challenging. Simulatability allows clinicians to audit predictions easily. They can manually inspect changes to outputs under slightly-perturbed inputs or identify when predictions are made due to systemic data bias rather than real causes. Our work shows that recurrent neural networks for predicting treatments for patients with sepsis or HIV can be trained to have simpler, tree-like decision boundaries with fewer than 25 nodes, while still predicting better than standalone trees.

Future work: interpretable representations of time-varying decisions. A key limitation of our tree-regularization work is that we interpret the predictions of a time-series model using data only from its latest time step. Improving simulatable explanations of time-series models will require careful thinking about better *representations* of complex time-varying trends (e.g. blood pressure was rising, but heart rate stabilized). Providing such explanations would help us understand the predictions of recurrent neural networks or BNP time-series models for many applications.

Research Vision: Machine Learning for Clinical Decision-Making.

In the next decade, I think answers to Q1, Q2, and Q3 can directly improve the daily decisions of clinicians treating individual patients and broaden scientific understanding of subtypes of depression and other diseases. My research agenda offers some crucial methods needed to answer these questions. I am excited to collaborate with others to achieve this goal, especially in making connections with reinforcement learning, natural language processing, probabilistic programming, and human-computer interaction. I'm particularly keen to work with UX designers and clinicians to collaboratively prototype the predictive analytics tools of the future so they actually improve care. Beyond strong methodological and clinical publications, I plan to deliver open-source tools to the broader ML community that make it possible to rapidly train, evaluate, inspect, and criticize a series of models to find the best approach for a given application. Finally, I am eager to integrate and deploy clinical decision support systems at the bedside, so patients can benefit from suggested treatments.

References

[1] M. R. Ackermann and J. Blömer. Bregman clustering for separable instances. In *Proceedings of the 12th Scandinavian conference on Algorithm Theory (SWAT)*, 2010.

[2] D. Arthur and S. Vassilvitskii. k-means++: The advantages of careful seeding. In *ACM-SIAM Symposium on Discrete Algorithms*, 2007.

[3] K. Ganchev, J. Graça, J. Gillenwater, and B. Taskar. Posterior regularization for structured latent variable models. *Journal of Machine Learning Research*, 11:2001–2049, Aug. 2010.

[4] M. Ghassemi, M. Wu, **M. C. Hughes**, et al. Predicting intervention onset in the ICU with switching state space models. In *AMIA Summit on Clinical Research Informatics*, 2017. [PDF].

[5] Y. Halpern, S. Horng, L. A. Nathanson, N. I. Shapiro, and D. Sontag. A comparison of dimensionality reduction techniques for unstructured clinical text. In *ICML Workshop on Clinical Data Analysis*, 2012.

[6] M. Hoffman, D. Blei, C. Wang, and J. Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14(1), 2013.

[7] G. Ji, **M. C. Hughes**, and E. B. Sudderth. From patches to images: A nonparametric generative model. In *International Conference on Machine Learning*, 2017. [PDF].

[8] M. J. Johnson, D. Duvenaud, A. B. Wiltschko, S. R. Datta, and R. P. Adams. Composing graphical models with neural networks for structured representations and fast inference. In *Neural Information Processing Systems*, 2016.

[9] D. Kingma and M. Welling. Auto-encoding variational Bayes. 2014.

[10] J. D. McAuliffe and D. M. Blei. Supervised topic models. In *Neural Information Processing Systems*, 2008.

[11] **M. C. Hughes**. BNPy: Bayesian nonparametric machine learning for Python, 2017. [PDF].

[12] **M. C. Hughes** and E. B. Sudderth. Memoized online variational inference for Dirichlet process mixture models. In *Neural Information Processing Systems*, 2013. [PDF].

[13] **M. C. Hughes** and E. B. Sudderth. Fast learning of clusters and topics via sparse posteriors. *ArXiv*, (1609.07521), 2016. [PDF].

[14] **M. C. Hughes**, L. Weiner, G. Hope, T. H. McCoy, R. H. Perlis, E. B. Sudderth, and F. Doshi-Velez. Prediction-constrained training for semi-supervised mixture and topic models. *ArXiv*, (1707.07341). [PDF].

[15] **M. C. Hughes**, D. I. Kim, and E. B. Sudderth. Reliable and scalable variational inference for the hierarchical Dirichlet process. In *Artificial Intelligence and Statistics*, 2015. [PDF].

[16] **M. C. Hughes**, W. Stephenson, and E. B. Sudderth. Scalable adaptation of state complexity for nonparametric hidden Markov models. In *Neural Information Processing Systems*, 2015. [PDF].

[17] **M. C. Hughes**, H. M. Elibol, T. H. McCoy, R. H. Perlis, and F. Doshi-Velez. Supervised topic models for clinical interpretability. In *NIPS Workshop on Machine Learning for Health*, 2016.

[18] **M. C. Hughes**, G. Hope, L. Weiner, T. McCoy, R. Perlis, E. Sudderth, and F. Doshi-Velez. Prediction-constrained topic models for antidepressant recommendation. In *NIPS Workshop on Machine Learning for Health*, 2017. [PDF].

[19] A. Mnih and K. Gregor. Neural variational inference and learning in belief networks. In *International Conference on Machine Learning*, 2014.

[20] A. S. Ross, **M. C. Hughes**, and F. Doshi-Velez. Right for the right reasons: Training differentiable models by constraining their explanations. In *International Joint Conference on Artificial Intelligence*, 2017. [PDF].

[21] M. Wu, **M. C. Hughes**, S. Parbhoo, M. Zazzi, V. Roth, and F. Doshi-Velez. Beyond sparsity: Tree regularization of deep models for interpretability. In *AAAI*, 2018. [PDF].

[22] C. Zhang and H. Kjellström. How to supervise topic models. In *ECCV Workshop on Graphical Models in Computer Vision*, 2014.

[23] J. Zhu, A. Ahmed, and E. P. Xing. MedLDA: maximum margin supervised topic models. *The Journal of Machine Learning Research*, 13(1):2237–2278, 2012.