Joshua V. Dillon

jvdillon@gmail.com

OBJECTIVE

Contribute to material advancements in machine learning and generative AI.

EDUCATION

Georgia Institute of Technology, Atlanta, GA 2009 – 2011
Ph.D., Computational Science & Engineering
- Thesis: "Stochastic m-Estimators for Controlling Accuracy-Cost Tradeoffs in Machine Learning"
- Advisor: Prof. Guy Lebanon
- GPA: 4.00/4.00
- (Thesis proves the asymptotic optimality of what would become known as BERT-style masking losses.)

Purdue University, West Lafayette, IN 2005 – 2008
M.S., Electrical & Computer Engineering
- GPA: 3.69/4.00

Michigan Technological University, Houghton, MI 2001 – 2005
B.S., Computer Engineering & Electrical Engineering (double major)
- Honors: Summa cum laude
- GPA: 3.90/4.00

EXPERIENCE

Luma, Staff Research Scientist, Foundational Models TL & Manager, Palo Alto, CA 2024 – 2025
- Led 6 person team which developed foundational (text,image,video)-to-video foundational models.
- Developed video compression autoencoders for generative video which significantly exceed SOTA.
- Architect of generative model, autoencoder, diffusion loss, and sampling code.
- Invented company-wide technique enabling complex, collaborative experimentation.

DeepMind, Staff Research Scientist, Video Foundations Team, Mountain View, CA 2024 – 2024
- Built prototype of Veo 1 (released at Google I/O, May 2024).

Google Research, Staff Research Scientist, CoreML Team, Mountain View, CA 2021 – 2023
- Led 3 person effort resulting in significant improvements to the autoencoder used by "VideoPoet." (Best paper; ICML'24.)
- Led 6 person effort resulting in SOTA self-supervised performance using novel ensembling.
- Co-developed backprop-like framework for analytically bounding Taylor series error.

Google Research, Senior Research Eng., BayesFlow Team, Mountain View, CA 2016 – 2021
- Architect and primary builder of TensorFlow Probability, the standard Python statistics toolbox for Jax & TensorFlow and basis for Pyro. (> 500k downloads / month.)
- Co-founded project (with Kevin Murphy and Rif A. Sauraus); now 13 researchers.
- Led numerous research efforts on uncertainty and approximate inference garnering 1000s of citations and dozens top ML publications (NeurIPS, UAI, AISTATS, ICLR, ICML, AABI, etc).
- Invited speaker at Google I/O 2019 (> 36k views.)
- Invited speaker at the 2019 TensorFlow Dev Summit. (> 59k views; most watched "non-core" talk.)
- Invited guest on "Coffee with a Googler." (> 20k views.)

Google Display Ads, Software Eng., Auction Team, Mountain View, CA 2011 – 2016

- Designed, implemented, theoretically justified, and launched a VCG-like auction which ultimately priced all display ads (all ads excluding those on google.com) from 2013–2019.
- Invented microeconomic statistical metrics which became standard launch criteria for all projects.
- Architect and TL for deep retrieval using ML to rank millions of entities with sub-millisecond latency.
- Co-founded (with Alex Smola) and built prototype of internal topic modeling project which ultimately generated O($Billion) in new net revenue.

OTHER RESEARCH EXPERIENCE

Research Assistant, Georgia Institute of Technology, Atlanta, GA 2009 – 2011

- Quantified the asymptotic accuracy of generative semi-supervised learning based on an extension of stochastic composite likelihood.
- Developed the stochastic m-estimator framework for controlling tradeoffs in machine learning such as computational cost, labeling cost, robustness. Examined large data statistical properties.
- Examined the effect of pseudo-periodicity in kernel density estimation and local likelihood methods and justified alternative techniques through analytical and empirical study.

Research Intern, Microsoft Research, Redmond, WA Summer 2009

- Developed a flexible optimization framework for constraining probabilistic models with imprecise domain knowledge. Applied this framework to find robust pseudo-relevance feedback models for IR.

Research Assistant, Purdue University, West Lafayette, IN 2006 – 2008

- Proposed a family of point estimators that resolve the computation-accuracy tradeoff present in maximum likelihood. Proved their consistency and provided formulas for asymptotic variance and computational complexity. Demonstrated usefulness for several graphical models.
- Developed the locally weighted bag of words framework for representing sequential text. Applied framework to several text analysis tasks: classification, segmentation, summarization, & visualization.
- Investigated machine translation and diffusion kernels for unsupervised metric learning of text.

Summer Scholar, Lawrence Livermore National Laboratory, Walnut Creek, CA Summer 2006

- Investigated unsupervised learning techniques for statistical process control. Applied these techniques to the Joint Genome Institute's DNA sequencing process to identify combinations of reagents, machines, and operators that lead to under-performing modes of operation.

Extreme Blue Intern, IBM, Austin, TX Summer 2005

- Developed unsupervised classification techniques which exploit harmonically related features. Applied this work for automatic error detection in the Linux kernel.

ENGINEERING EXPERIENCE

Intern, ThermoAnalytics Inc., Calumet, MI Spring 2005

- Solely designed and implemented a QScript-to-C translator optimized for numerical computing applications. Efforts included lexical analysis, context free grammar specification, developing an abstract syntax tree representation (code-emitting) routines.

Intern, IBM, Rochester, MN Summer 2004

- Implemented VHDL logic designs for the floating-point core of the Cell processor. Conducted timing analysis, synthesis, and testing of over 20 logic macros. Significantly improved team turnaround time by automating several report generating tasks (Perl) and implementing a tailored layout prototyping tool (Java).

Intern, IBM, Rochester, MN Summer 2003

- Designed, implemented, and packaged an SAP R/3 cluster management plug-in for iSeries Navigator (Java). Design goals included extensibility, graphical ease-of-use, and an aggressive release cycle to meet clients' demands. Efforts also involved the coordination of domestic and German colleagues. End product was delivered in a fully packaged form, ahead of schedule.

Intern, Michigan Department of Transportation, Cass City, MI Summer 2002

- Sole on-site inspector responsible for verifying contractors' adherence to design specifications. Responsible for chemical and physical quality control, logging payable items, and updating project plans.

TEACHING EXPERIENCE

Lecturer, Purdue University, West Lafayette, IN Spring & Fall 2008

- Two semester instructor for an ECE undergraduate course which acquaints students with scripted language software engineering tools, i.e., Python and korn shell. Responsibilities included curriculum design and delivering weekly lectures to 60+ students. One of only two grad student lecturers in dept.

Teaching Assistant, Purdue University, West Lafayette, IN Fall 2007

- Coordinating TA for "Microprocessor System Design and Interfacing," an undergraduate course which introduces microprocessor system design, assembly programming, and digital/analog interfaces. Held lab office hours and managed five undergraduate TAs.

FELLOWSHIPS

- Marshall Sherfield Postdoctoral Fellowship. Marshall Aid Commemoration Commission, 2011–13.
  - Full tuition and board for post-doctoral studies at any premier UK university.
  - Highly competitive: only two Americans selected per year.
  - Accepted by Zoubin Ghahramani to join Cambridge.
  - (Ultimately declined for personal reasons and instead joined Google.)
- DHS Fellowship in Data Analysis and Visual Analytics. Dept. of Homeland Security, 2010–11. (Full tuition and board at Georgia Tech.)
- Ross Graduate Fellowship. Purdue University, 2005–06. (Full tuition and board.)
- Board of Control Scholarship. Michigan Tech, 2001–05. (Full tuition.)

DISTINCTIONS

- US delegate, 57$^{\text{th}}$ Lindau Meeting of Nobel Laureates and Students. Germany, 2007.
- Eta Kappa Nu ECE Honor Society, Beta Chapter. Purdue University, 2006–08.
- Summa cum laude, Dept. of Electrical and Computer Engineering. Michigan Tech, 2005.
- Eta Kappa Nu ECE Honor Society, Beta Gamma Chapter. Michigan Tech, 2005.
- Phi Kappa Phi Honor Society. Michigan Tech, 2004.
- Tau Beta Pi Engineering Honor Society, Michigan Beta Chapter. Michigan Tech, 2004.
- Sheldon G. Hayes Foundation Scholar. Michigan Tech, 2003.
- Award of Excellence, Department of Mathematics. Michigan Tech, 2002.
- Dr. C. M. Carson Memorial Scholar. Michigan Tech, 2001.
- Michigan Merit Award. State of Michigan, 2001.
- Valedictorian, Cass City High School. Cass City, MI, 2001.

EXPERTISE

Machine Learning: deep learning, diffusion, energy model approximate inference, Probably approximately correct theory, graphical models, information theory, quantization.
Statistics: Bayesian methods, variational inference, expectation maximization, MCMC, Monte Carlo methods, point-/m-/z- estimators, large sample theory, stochastic processes.
Optimization: majorization minimization, trust region, convex optimization, online learning.
Engineering: numerical methods / stability, vector-processing (SIMD, AVX, GPU, TPU), multicore, cluster computing, API design, OSS development.
Applications: video, image, NLP, regression/classification, self-/un-/semi-/& supervised learning, federated learning, microeconomics.
Programming: Python, Jax, TensorFlow, NumPy, PyTorch, LaTeX, Go, C++, C, R, Matlab, Scheme, Perl, C, Bash, Linux.

SELECTED PUBLICATIONS

[1] G Comanici, ..., **JV Dillon**, ..., and NK Bhumihar, "Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities," 2025. arXiv: 2507.06261.

[2] D Kondratyuk, L Yu, X Gu, J Lezama, J Huang, R Hornung, H Adam, H Akbari, Y Alon, V Birodkar, Y Cheng, MC Chiu, **JV Dillon**, I Essa, A Gupta, M Hahn, A Hauth, D Hendon, A Martinez, D Minnen, D Ross, G Schindler, M Sirotenko, K Sohn, K Somandepalli, H Wang, J Yan, MH Yang, X Yang, B Seybold, and L Jiang, "Videopoet: A large language model for zero-shot video generation," in *ICML*, 2024. arXiv: 2312.14125.

[3] S Abu-El-Haija, **JV Dillon**, B Fatemi, K Axiotis, N Bulut, J Gasteiger, B Perozzi, and M Bateni, "Submix: Learning to mix graph sampling heuristics," in *UAI*, 2023. arXiv: 2312.14125. [Online]. Available: https://proceedings.mlr.press/v216/abu-el-haija23a/abu-el-haija23a.pdf.

[4] Y Ruan, S Singh, W Morningstar, AA Alemi, S Ioffe, I Fischer, and **JV Dillon**, "Weighted ensemble self-supervised learning," in *ICLR*, 2023. arXiv: 2211.09981. [Online]. Available: https://openreview.net/forum?id=CL-sVR9pvF.

[5] M Streeter and **JV Dillon**, "Sharp taylor polynomial enclosures in one dimension," 2023. arXiv: 2308.00679.

[6] E Vedadi, **JV Dillon**, PA Mansfield, K Singhal, A Afkanpour, and WR Morningstar, "Federated variational inference: Towards improved personalization and generalization," 2023. arXiv: 2305.13672.

[7] W Morningstar, AA Alemi, and **JV Dillon**, "Pac^m-Bayes: Narrowing the empirical risk gap in the misspecified Bayesian regime," in *AISTATS*, vol. 151, 2022, pp. 8270–8298. arXiv: 2010.09629. [Online]. Available: https://proceedings.mlr.press/v151/morningstar22a.html.

[8] M Streeter and **JV Dillon**, "Automatically bounding the Taylor remainder series: Tighter bounds and new applications," 2022. arXiv: 2212.11429.

[9] AA Alemi, W Morningstar, B Poole, I Fischer, and **JV Dillon**, "VIB is half Bayes," 2021. arXiv: 2011.08711. [Online]. Available: https://openreview.net/forum?id=97FiVYw4mrF.

[10] W Morningstar, S Vikram, C Ham, A Gallagher, and **JV Dillon**, "Automatic differentiation variational inference with mixtures," in *AISTATS*, 2021, pp. 3250–3258. arXiv: 2003.01687. [Online]. Available: https://proceedings.mlr.press/v130/morningstar21b.html.

[11] J Lao, C Suter, I Langmore, C Chimisov, A Saxena, P Sountsov, D Moore, RA Saurous, MD Hoffman, and **JV Dillon**, "tfp.mcmc: Modern Markov chain Monte Carlo tools built for modern hardware," in *PROBPROG*, 2020. arXiv: 2002.01184.

[12] W Morningstar, C Ham, AG Gallagher, B Lakshminarayanan, AA Alemi, and **JV Dillon**, "Density of states estimation for out-of-distribution detection," in *AISTATS*, 2020. arXiv: 2006.09273. [Online]. Available: https://proceedings.mlr.press/v130/morningstar21a.html.

[13] D Piponi, D Moore, and **JV Dillon**, "Joint distributions for TensorFlow Probability," in *PROBPROG*, 2020. arXiv: 2001.11819.

[14] J Swiatkowski, K Roth, B Veeling, L Tran, **JV Dillon**, J Snoek, S Mandt, T Salimans, R Jenatton, and S Nowozin, "The k-tied normal distribution: A compact parameterization of Gaussian mean field posteriors in Bayesian neural networks," in *ICML*, vol. 119, 2020, pp. 9289–9299. arXiv: 2002.02655. [Online]. Available: https://proceedings.mlr.press/v119/swiatkowski20a.html.

[15] L Tran, BS Veeling, K Roth, J Swiatkowski, **JV Dillon**, J Snoek, S Mandt, T Salimans, S Nowozin, and R Jenatton, "Hydra: Preserving ensemble diversity for model distillation," in *ICML Uncertainty & Robustness in Deep Learning Workshop*, 2020. arXiv: 2001.04694. [Online]. Available: http://www.gatsby.ucl.ac.uk/~balaji/udl2020/accepted-papers/UDL2020-paper-026.pdf.

[16] Y Ovadia, E Fertig, J Ren, Z Nado, D Sculley, S Nowozin, **JV Dillon**, B Lakshminarayanan, and J Snoek, "Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift," in *NeurIPS*, vol. 32, 2019. arXiv: 1906.02530. [Online]. Available: https://proceedings.neurips.cc/paper/2019/hash/8558cb408c1d76621371888657d2eb1d-Abstract.html.

[17] J Ren, PJ Liu, E Fertig, J Snoek, R Poplin, MA DePristo, **JV Dillon**, and B Lakshminarayanan, "Likelihood ratios for out-of-distribution detection," in *NeurIPS*, 2019. arXiv: 1906.02845. [Online]. Available: https://proceedings.neurips.cc/paper/2019/hash/1e79596878b2320cac26dd792a6c51c9-Abstract.html.

[18] AA Alemi, I Fischer, and **JV Dillon**, "Uncertainty in the variational information bottleneck," in *UAI Uncertainty in Deep Learning Workshop*, 2018. arXiv: 1807.00906. [Online]. Available: http://www.gatsby.ucl.ac.uk/~balaji/udl-camera-ready/UDL-18.pdf.

[19] AA Alemi, B Poole, I Fischer, **JV Dillon**, RA Saurous, and K Murphy, "Fixing a broken ELBO," in *ICML*, 2018, pp. 159–168. arXiv: 1711.00464. [Online]. Available: https://proceedings.mlr.press/v80/alemi18a.html.

[20] MD Hoffman, P Sountsov, **JV Dillon**, I Langmore, D Tran, and S Vasudevan, "NeuTra-lizing bad geometry in Hamiltonian Monte Carlo using neural transport," in *AABI*, 2018. arXiv: 1903.03704. [Online]. Available: http://approximateinference.org/2018/accepted/HoffmanEtAl2018.pdf.

[21] **JV Dillon** and I Langmore, "Quadrature compound: An approximating family of distributions," 2018. arXiv: 1801.03080.

[22] AA Alemi, I Fischer, **JV Dillon**, and K Murphy, "Deep variational information bottleneck," in *ICLR*, 2017. arXiv: 1612.00410. [Online]. Available: https://openreview.net/forum?id=HyxQzBceg.

[23] **JV Dillon**, I Langmore, D Tran, E Brevdo, S Vasudevan, D Moore, B Patton, A Alemi, MD Hoffman, and RA Saurous, "TensorFlow Distributions," in *The 45th ACM SIGPLAN Symposium on Principles of Programming Languages*, 2017. arXiv: 1711.10604. [Online]. Available: https://popl18.sigplan.org/details/pps-2018/8/TensorFlow-Distributions.

[24] **JV Dillon**, "Stochastic m-estimators: Controlling accuracy-cost tradeoffs in machine learning," Ph.D. dissertation, Georgia Institute of Technology, 2011. [Online]. Available: https://smartech.gatech.edu/handle/1853/42913.

[25] K Collins-Thompson and **JV Dillon**, "Controlling the search for expanded query representations by constrained optimization in latent variable space," in *SIGIR Workshop on Query Representation and Understanding*, 2010. [Online]. Available: http://www-personal.umich.edu/~kevynct/pubs/sigir10-queryws.pdf.

[26] **JV Dillon**, K Balasubramanian, and G Lebanon, "Asymptotic analysis of generative semi-supervised learning," in *ICML*, 2010. arXiv: 1003.0024. [Online]. Available: https://icml.cc/Conferences/2010/papers/107.pdf.

[27] **JV Dillon** and K Collins-Thompson, "A unified optimization framework for finding reliable pseudo-relevance feedback models," in *CIKM*, 2010. [Online]. Available: https://dl.acm.org/doi/10.1145/1871437.1871573.

[28] **JV Dillon** and G Lebanon, "Stochastic composite likelihood," in *JMLR*, vol. 11, 2010, pp. 2597–2633. [Online]. Available: https://www.jmlr.org/papers/v11/dillon10a.html.

[29] S Kim, **JV Dillon**, and G Lebanon, "Cumulative revision map," 2010. arXiv: 1205.3205.

[30] **JV Dillon** and G Lebanon, "Statistical and computational tradeoffs in stochastic composite likelihood," in *AISTATS*, 2009. arXiv: 1003.0691. [Online]. Available: https://proceedings.mlr.press/v5/dillon09a.html.

[31] **JV Dillon**, Y Mao, G Lebanon, and J Zhang, "Statistical translation, heat kernels, and expected distances," in *UAI*, 2007, pp. 93–100. arXiv: 1206.5248. [Online]. Available: http://bengio.abracadoudou.com/lce/papers/12.pdf.

[32] G Lebanon, Y Mao, and **JV Dillon**, "The locally weighted bag of words framework for documents," in *JMLR*, vol. 8, 2007, pp. 2405–2441. [Online]. Available: https://jmlr.org/papers/v8/lebanon07a.html.

[33] Y Mao, **JV Dillon**, and G Lebanon, "Sequential document visualization," in *IEEE Transactions on Visualization and Computer Graphics*, vol. 13, 2007, pp. 1208–1215. [Online]. Available: http://theanalysisofdata.com/gl/tvcg07.pdf.