Michael Munn, PhD

New York, NY

mikemunn@gmail.com | 646.409.1398
linkedin.com/in/munnm | Google scholar

Experienced Research Software Engineer with a demonstrated history of success in scoping, developing, and productionzing large scale machine learning solutions across multiple verticals. Highly skilled in deep learning, artificial intelligence, and Cloud computing with a mathematics PhD, postdoc and academic experience. Published author of multiple books, research articles, and blog posts.

Skills

Programming: Python, SQL, bash, R, C++
Tools: Google Cloud Platform, Tensorflow, PyTorch, JAX, Kubeflow, BigQuery, Dataflow, Scikit-Learn, Jupyter, Flask, Explainability methods, supervised and unsupervised techniques, optimal transportation
Publications/Presentations: Machine Learning Design Patterns (O'Reilly '20), Explainable AI for Practitioners (O'Reilly '22), 50+ presentations at seminars/conferences/colloquia for technical and non-technical audiences and peer-reviewed publications

Experience

Google Research, Senior Software Engineer **Feb/2022 – present**
I work within Google's Athena Research group focused on creating useful solutions to fundamental problems in theory and algorithms, large-scale machine learning, large language models, and speech with direct impact on Google's product. I have a strong contribution to research communities within and outside of Google and my research has appeared in top tier conferences.

Google Cloud, ML Solutions Engineer **Apr/2018 – Feb/2022**
My role responsibilities were split between Google Cloud's AI Services and the Advanced Solutions Lab. Within AI Services, we built and productionized Cloud AI solutions for customers across all verticals. The Advanced Solutions Lab provided dedicated ML training on best practices for developing end-to-end machine learning solutions.

- **Tech Lead**, Google.org Fellowship with The Trevor Project
  - developed and deployed a custom conversation simulator model using PyTorch and GPT2
  - led a team of six ML SWEs across Google and Trevor
  - determined scope for the 6 month project from exploratory ML development to a deployed model currently in production and being used by Trevor trainees
  - In 2021, trainees used our conversation simulator for 4,919 hrs of training, in essence removing 35-50 role-play shifts for trainers and the potential to graduate an additional 200-300 new counselors
  - regularly coordinated with business stakeholders with Trevor and Google to align on delivery and key objectives to identify and resolve blockers
  - external press: MIT Tech Review, Time Magazine's Best Inventions of 2021, Google.org blog post, Today show segment, and Google I/O '21
- **Engineering and technical Lead** for multiple ML engagements with Google Cloud customers across various verticals delivering end-to-end ML solutions from development to productionisation
  - Led scoping and ML/AI Feasibility engagements with Cloud customers & AI Accelerator grantees to assess use cases and shape delivery
- **Delivered dedicated ML Instruction** for Google's Advanced Solutions Lab, an intensive, immersive ML training course combining Advanced ML curriculum with Tensorflow/GCP and sprint-based Open Project work.
  - consistently received 95+% Overall CSAT and NPS, and 100% Trainer Communications and 100% Trainer Technical skill on customer feedback
  - developed and maintained the external code base (on github) consisting of hands-on labs used during training delivery

Accenture, Data Science Consultant **Jan/2017 – Apr/2018**

• Developed DNN and custom logistic Estimator in Tensorflow to predict future server failure, implemented LIME framework and other explainability methods to provide insight to root cause analysis of failure
• Created a multi-class Random Forest model in sk-learn to classify aircrafts in flight across 5 separate labels, applied SMOTE to handle class imbalance, implemented end-to-end pipeline for final RNN model handling data ingestion through BigQuery, preprocessing via Dataflow, modeling in Tensorflow and accuracy analysis
• Built a LSTM model and Tensor2Tensor model in Tensorflow to provide route prediction of aircraft flight data
• Applied data driven analytics and ML modeling to recognize $8MM in reducible cost for one of largest utility companies in the US
• Built robust modeling pipelines in python for supervised (regularized regression, boosting, random forest) and unsupervised (k-means) learning, cutting down analysis time from days to hours

Insight Health Data Science Fellow **Sept/2016 – Dec/2016**

• Acquired user FitBit data analyzed sleep and activity data per user using python/pandas and mixed effect models in R
• Developed and fine-tuned multiple ML models to provide a personalized user sleep recommendation, based on their personal data and crowdsourced data
• Deployed an interactive webapp on AWS using Flask and Bootstrap in python to serve the app

(prior academic work experience)
New York University, Courant Institute, Clinical Assistant Professor **2014 – 2016**
University of Missouri, Assistant Professor **2011 – 2014**
University of Warwick, National Science Foundation Postdoctoral Fellow **2009 – 2011**
CUNY, NYCCT, Assistant Professor **2008 – 2011**

Education

City University of New York, NY, NY
Ph.D in Mathematics (focus area: geometric analysis and topology) **2008**
M.Phil in Mathematics **2006**
University of Notre Dame, South Bend, IN
B.S. in Honors Mathematics **2001**

Publications

Books

- Explainable AI for Practitioners, O'Reilly, November 2022
  (with D. Pitman) This book is a collection of the most effective and commonly used techniques for explaining why an ML model makes the predictions it does. We discuss the many aspects of Explainable AI including the challenges, metrics for success, caveats and best practices.
- Machine Learning Design Patterns, O'Reilly, November 2020
  (with V. Lakshmanan and S. Robinson) Design patterns are formalized best practices that are used to solve common problems when designing an application or system. This book captures best practices and solutions to recurring problems in machine learning, from data exploration, model building and MLOps.

Recent Research Papers

- A margin-based multiclass generalization bound via geometric complexity, ICML 2023, Topology, Algebra and Geometry Workshop (with B. Dherin, X. Gonzalvo). We derive a new upper bound on the generalization error which scales with the margin-normalized geometric complexity of the network and which holds for a broad family of data distributions and model classes.
- Why neural networks find simple solutions: the many regularizers of geometric complexity, NeurIPs 2022
  (with B. Dherin, M. Rosca, D. Barrett) We introduce a new measure of model complexity, called Geometric Complexity that has properties suitable for the analysis of deep neural networks. We use theoretical and empirical techniques to demonstrate that many commonly used training heuristics all act to control geometric complexity. We argue that the geometric complexity provides a convenient proxy for neural

network performance.

- The Geometric Occam's Razor Implicit in Deep Learning, NeurIPS 2021, Workshop on Optimization
  (with B. Dherin, D. Barrett) We argue that over-parameterized neural networks trained with stochastic gradient descent are subject to a Geometric Occam's Razor; that is, these networks are implicitly regularized by the geometric model complexity. We prove how this geometric pressure arises naturally during gradient descent and point to methods which can induce learned functions that are robust to input perturbations.
- COT-GAN: Generating Sequential Data via Causal Optimal Transport, NeurIPS 2020
  (with T. Xu, L. Wenliang, B. Acciaio) We develop an adversarial algorithm to train implicit generative models for producing sequential data. The loss function of this algorithm is formulated using ideas from Causal Optimal Transport (COT), which combines classic optimal transport methods with an additional temporal causality constraint. We include an entropic penalization term which allows for the use of the Sinkhorn algorithm when computing the optimal transport cost. Our experiments show effectiveness and stability of COT-GAN when generating both low- and high-dimensional time-series data. The success of the algorithm also relies on a new, improved version of the Sinkhorn divergence which demonstrates less bias in learning.

Pre-Google

(w/ S. Lakzian) On the Size of a Ricci Flow Neckpinch via Optimal Transport - Analysis and Geometry of Metric Measure Spaces, 2021

(w/ Q. Deng, F. Galaz-Garcia, L. Guijarro) Three-Dimensional Alexandrov spaces with positive or nonnegative Ricci curvature, Potential Analysis, 2017

(w/ L. Bandara, S. Lakzian) Geometric singularities and a flow tangent to the Ricci flow, Annali della S.N.S di Pisa, 2015

Alexandrov spaces with large volume growth, Journal of Mathematical Analysis and Applications, 2015

(w/ S. Lakzian) Super Ricci flow for disjoint unions, Analysis and Geometry of Metric Measure Spaces, 2012

Volume growth and the topology of pointed Gromov-Hausdorff limits, Differential Geometry and Its Applications, 2010

Volume growth and the topology of manifolds with nonnegative Ricci curvature, Journal of Geometric Analysis, 2010

(w/ D. Garbin, J. Jorgenson) On the appearance of Eisenstein series through degeneration, Commentarii Mathematici Helvetici, 2008

Blog Posts

How to deploy interpretable models on Google Cloud Platform, Toward Data Science, 2020, github code

Building a document understanding pipeline with Google Cloud, Google Cloud AI blog, 2019, github code