MICAH J SMITH

OpenEvidence
Massachusetts Institute of Technology

245 Main St, Floor 2
Cambridge, MA 02142
+1 805 657-2325
micahs@alum.mit.edu
www.micahsmith.com
he/him

RESEARCH INTERESTS: machine learning systems, human-computer interaction, open-source, software engineering, crowd computing, databases, big data computing

EDUCATION

Massachusetts Institute of Technology, Department of EECS 2016–2021
Ph.D., Computer Science
Thesis: Collaborative, Open, and Automated Data Science
Adviser: Dr. Kalyan Veeramachaneni

Massachusetts Institute of Technology, Department of EECS 2016–2018
S.M., Computer Science
Thesis: Scaling Collaborative Open Data Science
Adviser: Dr. Kalyan Veeramachaneni

Columbia University, Columbia College 2010–2014
B.A., Economics-Mathematics, *cum laude*

RESEARCH EXPERIENCE

MIT LIDS, Data to AI Lab — Graduate Research Assistant 2016–2021
- Researching and developing algorithms and systems to support data science and machine learning teams by focusing on collaboration and open-source development. Projects include Ballet, a lightweight software framework for collaborative feature engineering in an open-source setting, Assemblé, a development environment for collaborative data science, and FeatureHub, a cloud system for collaborative feature engineering.
- Researching and developing automated machine learning (AutoML) systems to make data science easier and more efficient. Projects include ML Bazaar, an ecosystem for composable, end-to-end machine learning, AutoBazaar, a general-purpose AutoML system, and ATMSeer, an interactive visualization frontend for AutoML systems.
- Researched applications of query optimization for imputation in a database system for early dataset exploration. Projects include ImputeDB, a prototype database engine that dynamically imputes missing values during database query execution.
- *Adviser*: Kalyan Veeramachaneni

Federal Reserve Bank of New York, Research Group — Senior Research Analyst 2014–2016
- *DSGE Group*. Lead research analyst and developer of FRBNY DSGE model MATLAB codebase and open-source, high-performance Julia implementation. Developed optimal monetary policy models, alternative monetary policy experiments, macroeconomic forecasts, and other policy analysis.
- *Policy Reaction Functions.* Developed and implemented MATLAB model to analyze policy reaction functions of FOMC members and public forecasters. Addressed zero lower bound problem by treating target interest rate as latent variable and applying Bayesian MCMC methods.
- Other projects included analysis of consumer housing expectations housing dataset, mortgage interest deduction perceptions survey experiment, visualization and analysis of NYC subprime mortgage delinquency flows.

· *Advisers*: Marco Del Negro, Marc Giannoni

Columbia University, Department of Mathematics — Undergraduate Research Fellow 2013
· Led team in Python modeling of novel elliptic curve arithmetic and cryptographic algorithms to explore ideas for performance improvements, by extending one approach to Miller's algorithm for computing the Weil pairing.
· *Advisers*: Mike Woodbury, Marc Masdeu

PUBLICATIONS

CHIL '23 — E. Lehman, E. Hernandez, D. Mahajan, J. Wulff, **M. Smith**, Z. Ziegler, D. Nadler, P. Szolovits, A. Johnson, E. Alsentzer. "Do We Still Need Clinical Language Models?" *Proceedings of the Conference on Health, Inference, and Learning, PMLR 209:578-597.* 2023.

PhD Thesis '21 — **M. Smith**. "Collaborative, Open, and Automated Data Science." *MIT Ph.D. Thesis.* 2021.

CSCW '21 — **M. Smith**, J. Cito, K. Lu, and K. Veeramachaneni. "Enabling Collaborative Data Science Development with the Ballet Framework." *Proceedings of the ACM on Human-Computer Interaction (CSCW)*. 2021.

CSUR '21 — S. Santu, M. Hassan, **M. Smith**, L. Xu, C. Zhai, K. Veeramachaneni. "AutoML to Date and Beyond: Challenges and Opportunities." To appear in *ACM Computing Surveys.* 2021.

Preprint '21 — **M. Smith**, J. Cito, and K. Veeramachaneni. "Meeting in the Notebook: A Notebook-Based Environment for Micro-Submissions in Data Science Collaborations." *arXiv preprint 2103.15787.* 2021.

SIGMOD '20 — **M. Smith**, C. Sala, J.M. Kanter, K. Veeramachaneni. "The Machine Learning Bazaar: Harnessing the ML Ecosystem for Effective System Development." *Proceedings of the 2020 International Conference on Management of Data (SIGMOD)*. 2020.

CHI '20 — D. Liu*, **M. Smith***, K. Veeramachaneni. "Understanding User-Bot Interactions for Small-Scale Automation in Open-Source Development." *CHI Conference on Human Factors in Computing Systems Extended Abstracts (CHI)*. 2020. (*equal contribution)

MLSys '20 — **M. Smith**, K. Lu, K. Veeramachaneni. "Demonstration of Ballet: A Framework for Open-source Collaborative Feature Engineering." In *Third Conference on Machine Learning and Systems (MLSys)*. 2020.

CHI '19 — Q. Wang, Y. Ming, Z. Jin, Q. Shen, D. Liu, **M. Smith**, K. Veeramachaneni, H. Qu. "ATMSeer: Increasing Transparency and Controllability in Automated Machine Learning." In *CHI Conference on Human Factors in Computing Systems Proceedings (CHI)*. 2019.

NeuRIPS '18 — **M. Smith,** K. Lu, and K. Veeramachaneni. "Ballet: A lightweight framework for open-source, collaborative feature engineering." *Workshop on Systems for ML and Open Source Software at NeuRIPS 2018*. 2018.

S.M. Thesis '18 — **M. Smith**. "Scaling Collaborative Open Data Science." *MIT S.M. Thesis.* 2018.

DSAA '17 — **M. Smith**, R. Wedge, and K. Veeramachaneni. "FeatureHub: towards collaborative data science." *IEEE International Conference on Data Science and Advanced Analytics (DSAA).* 2017.

VLDB ‘17 J. Cambronero*, J. Feser*, **M. Smith***, and S. Madden. “Query optimization for dynamic imputation.” *Proceedings of the VLDB Endowment (VLDB).* 2017.

MISCELLANY

- M. Del Negro, G. Eggertsson, A. Ferrero, and N. Kiyotaki. “The Great Escape? A Quantitative Evaluation of the Fed's Liquidity Facilities.” *American Economic Review*, 107(3): 824-57.` 2017. (Substantial contribution)
- M. Del Negro, M. Giannoni, and **M. Smith**. “The Macro Effects of the Recent Swing in Financial Conditions.” *Liberty Street Economics.* May 25, 2016.
- M. Del Negro, M. Giannoni, P. Li, E. Moszkowski, and **M. Smith**. “The FRBNY DSGE Model Meets Julia.” *Liberty Street Economics*. December 3, 2015.
- Z. Cranko, P. Li, S. Lyon, E. Moszkowski, **M. Smith**, and P. Winant. “The DSGE MATLAB to Julia Transition: Improvements and Challenges.” December 3, 2015.
- M. Del Negro, M. Giannoni, E. Moszkowski, S. Shahanaghi, and **M. Smith**. “The FRBNY DSGE Model Forecast - November 2015.” *Liberty Street Economics.* December 1, 2015.
- M. Del Negro, M. Giannoni, and C. Patterson. “The forward guidance puzzle.” *Federal Reserve Bank of New York Staff Reports.* October 2012, revised December 1, 2015. (Substantial contribution)
- A. Fuster, B. Zafar, and **M. Smith**. “Just Released: 2015 SCE Housing Survey Shows Households Optimistic about Housing Market.” *Liberty Street Economics.* May 28, 2015.
- A. Fuster, B. Zafar, and **M. Smith**. “Survey of Consumer Expectations: Housing Survey - 2015: Report.” *New York Fed Microeconomics.* May 28, 2015.
- M. Del Negro, M. Giannoni, M. Cocci, S. Shahanaghi, and **M. Smith**. “Why are Interest Rates So Low?” *Liberty Street Economics*. May 20, 2015.
- M. Del Negro, M. Giannoni, M. Cocci, S. Shahanaghi, and **M. Smith**. “The FRBNY DSGE Model Forecast - April 2015.” *Liberty Street Economics.* May 18, 2015.
- M. Del Negro and **M. Smith**. Discussion of “Macroeconomic Effects of the Federal Reserve's Unconventional Monetary Policies.” *Banque de France and Federal Reserve Bank of New York Workshop on Forward Guidance and Expectations.* May 1, 2015.

POSTERS

- “Demonstration of Ballet: A Framework for Open-Source Collaborative Feature Engineering.” *MLSys*. March 2, 2020.
- “Ballet: A lightweight framework for open-source, collaborative feature engineering.” *Workshop on Systems for Machine Learning and Open Source Software at NeuRIPS*. December 7, 2018.
- “Query optimization for dynamic imputation.” *VLDB*. August 30, 2017.
- “FeatureHub: a cloud platform for feature engineering.” *SDSCon*. April 21, 2017.

TALKS

- “Enabling Collaborative Data Science Development with the Ballet Framework.” *CSCW*. October 26, 2021.
- “A New Approach to Collaborative Data Science with the Ballet Framework.” *LIDS Student Conference*. February 3, 2021.
- “The Machine Learning Bazaar: Harnessing the ML Ecosystem for Effective System Development.” *SIGMOD*. June 16, 2020.
- “FeatureHub: towards collaborative data science.” *DSAA*. October 20, 2017.

PROFESSIONAL EXPERIENCE

OpenEvidence (Cambridge, MA) – Infrastructure Lead/Machine Learning Engineer 2022-pres.

- I own backend, site reliability, DevOps, infra, security, compliance, on-call (GCP, Django, Postgres, Python, GitHub Actions, Terraform, HIPAA) as the Technical Lead of the Core Engineering team
- Leading ML infrastructure efforts for finetuning and inferencing large language models

Twitter (Boston, MA) — Machine Learning Engineer 2021–2022

- Developed machine learning platform frameworks and infrastructure on the ML Pipelines team (TFX, Kubeflow Pipelines, TensorFlow, Apache Beam, GCP BigQuery, GCP Dataflow)

Botkeeper (Boston, MA) — Machine Learning Engineer (part-time) 2019–2021
- Lead company-wide ML engineering efforts in transaction classification for accounting
- Designed and implemented automatic model retraining, ML metrics collection, storage, and querying (Python, MongoDB, Kubernetes)
- Developed models to address class imbalance and distribution shift (scikit-learn, Keras)
- Researched and developed prototype of "global model" to transfer contextual information about transactions across clients using deep character-level embedding module (TensorFlow, Keras)

Twitter (New York, NY) — Machine Learning Engineering Intern (Cortex Machine Learning) 2018
- Re-designed internal machine learning workflow execution system for easy-to-use, automated hyper parameter tuning functionality (Python, Airflow).
- Integrated workflow system with internal Bayesian optimization hyper parameter tuning service (Python, Whetlab/Spearmint, Django).
- Enabled one-line smart model tuning, used for production models within days.

Kensho Technologies (Cambridge, MA) — Machine Learning Intern 2017
- Developed machine learning model to predict trading behaviors at Treasuries desk of major US dealer.

Carlisle Development Group (Miami, FL) — Summer Associate 2012
- Overhauled in-house hard cost estimating with ground-up development of new econometric model.

Pacific Coast Business Times (Santa Barbara, CA) — Finance/Editorial Intern 2011
- Composed special features about regional economic players, market analysis, and opinion pieces.

TEACHING EXPERIENCE

Graduate Teaching Assistant — MIT 6.031 Software Construction Spring 2020
- Held lab hours, critiqued design of and graded programming projects, managed course calendars
- Helped plan and implement rapid transition to remote learning during emergency move off campus due to COVID-19 pandemic

MENTORING

- Gannon Barnett (Engineering Intern, Summer 2020). Supervised project in structured data extraction from PDF documents and provided feedback on design and implementation.
- Kelvin Lu (M.Eng. 2018–2019). Supervised 9 month research program in collaborative data science culminating in co-authored papers and thesis, "Feature engineering and evaluation in lightweight systems."
- Erica Moszkowski, Pearl Li (Research Analysts, 2015–2016). As Senior RA in a research group, mentored two junior RAs in development of a Julia codebase for economic policy analysis.

OTHER ACTIVITIES

- Organizer/Mentor, MIT EECS Graduate Application Assistance Program (2020 - pres.)
- Chair, Pub Night, MIT EECS Graduate Student Association (2019)
- Chair, Coffee Hour, MIT EECS Graduate Student Association (2018–2019)
- Organizer, MIT DAI Lab Data Science Reading Group (2017–2021)
- VP Communications, MIT EECS Graduate Student Association (2018)
- Co-chair, LIDS Student Conference (2017–2018)
- Bartender, Muddy Charles Pub (2017–2019)
- Open source developer (BTB, ATM, Julia, Mocha.jl, Airflow, jupyterhub-client, pelican-plugins, DSGE.jl, FredData.jl, BlsData.jl, etc.)
- Columbia University Tennis Club, President (2010–2014)
- Moneythink Columbia, Mentor and Executive Board member (2011–2012)
- Columbia Daily Spectator, Copy Editor (2010–2012)
- Running, biking, tennis, basketball, reading, coffee, crosswords, chess, playing with my dog Mamba

PRESS

- MIT News. “Mentorship program encourages underrepresented graduate applicants in EECS.” October 13, 2020. https://news.mit.edu/2020/gaap-mentorship-encourages-underrepresented-students-eecs-1013
- MIT EECS. “Closing the GAAP: a new mentorship program encourages underrepresented students in the final stretch of their academic marathon.” September 29, 2020. https://www.eecs.mit.edu/news-events/media/closing-gaap-new-mentorship-program-encourages-underrepresented-students-final
- EECS THRIVE. “MIT students launch ambitious mentorship program for graduate admissions in electrical engineering and computer science.” September 15, 2020. https://www.thrive-eecs.mit.edu/news-eecs-gaap-launch
- The Register. “MIT boffins hope to speed up analytics with GitHub-style platform.” November 3, 2017. https://www.theregister.com/2017/11/03/crowdsourcing_for_data_analysis_mit_boffins_propose_github_for_feature/
- MIT News. “Crowdsourcing big-data analysis.” October 30, 2017. https://news.mit.edu/2017/crowdsourcing-big-data-analysis-1030

AWARDS

FRBNY Performance Excellence Awards 2014–2015
DSGE Model Development, SCE Housing Report, Automated Chart Creation and Repository Development, Launching “US Economy in a Snapshot,” DSGE.jl Leadership and Contributions
Columbia College Dean’s List 2010–2014
Fairburn Memorial Scholar, National Merit Scholar 2010

SKILLS

General	Python, TypeScript, JavaScript, Java, Bash, Julia, C++, C, Scala, MATLAB, Haskell
Data science/ machine learning	tensorflow, TFX, Kubeflow, PyData (pandas, numpy, scikit-learn, seaborn, matplotlib), Julia, MATLAB
Data	MongoDB, SQL, Airflow, SQLAlchemy, Spark, Dask, BigQuery, Beam, Dataflow
DevOps	Docker, Docker Compose, Kubernetes, Travis, AWS (EC2, S3, EKS, etc.), Heroku, GH Apps
Web	Flask, pelican, jinja2, node, jQuery, Google Apps Script, Django, HTML/CSS, React, Tornado
Tools	git, GitHub, GitLab, Jupyter Lab/Notebook/Hub, vim, *nix, LaTeX, make, sphinx, VS Code, PyCharm, Eclipse, Bazel, invoke
Foreign languages	Italian (Intermediate), Spanish (Intermediate)

REFERENCES

Advisors and collaborators who have written references

- Kalyan Veeramachaneni, Principal Research Scientist at MIT
- Sam Madden, Professor at MIT
- Marco Del Negro, Vice President at Federal Reserve Bank of New York
- Marc Giannoni, Research Director at Federal Reserve Bank of Dallas
- Seyhan Erden, Professor at Columbia University
- Mike Woodbury, Professor at University of Cologne

The latest version of this document is available at www.micahsmith.com/files/cv.pdf