Natalie **Durgin**

Mathematician · Data Scientist

www.nataliedurgin.com | nataliedurgin | nataliedurgin

Education

Rice University — *Houston, TX*

Ph.D. in Mathematics — *2009-2015*

- Thesis: Geometric Invariant Theory Quotient of the Hilbert Scheme of Six Points on the Projective Plane
- Advisor: Brendan Hassett

Harvey Mudd College — *Claremont, CA*

B.S. in Mathematics, Presidential Scholar — *2005-2009*

- Thesis: The Abelian Sandpile Model on Symmetric Graphs
- Advisor: Francis Su

Skills

Programming	Python, Scala, SQL, LaTeX, (Mac OSX, Ubuntu, Centos, Docker, AWS S3/Athena/EMR/ECR/ECS)
Data Tools	Spark(ML, Mllib, GraphX), Scikit-learn, Pandas, MatplotLib, Jupyter, Gensim, Presto, Hive/Hadoop, Luigi, Airflow
Data Science	Association Rules, (Un)Supervised Learning, Random Forest, NLP, Word2Vec, Fasttext, Neural Nets
Mathematics	Compressed Sensing, Anomaly Detection, Algebraic Geometry, Geometric Invariant Theory

Experience

Spiceworks — *Austin, TX*

Senior Data Scientist — *April 2017 - PRESENT*

- *Community Question Answering:* Developed algorithm to programmatically cluster similar questions from a noisy corpus based on semantic similarities and not merely lexical ones. Achieved substantial lift in site-wide engagement metrics. (gensim word2vec, stack-overflow corpora, following Charlet and Damnati, 2017)
- *On-Network Ad-serving Strategy:* Developed offline simulation strategy to pit DFP's performance against the performance of a custom ad-serving algorithm that leveraged internal data. (Pyspark ML Random Forest Regressor, AWS S3/EMR)
- *Email Delivery Optimization:* Sought to find an email delivery policy that optimizes for the lifetime value of users (reinforcement learning, ultimately scrapped in favor of more expedient approach)

Data Scientist — *June 2015 - April 2017*

- *Neuro-Marketing:* Scanned the brains of over 40 volunteers to collect EEG data and search for patterns in their reactions to viewing various types of advertising (OpenBCI, Spiceworks internal hackathon)
- *List-Cutting-Tool:* Trained random forest to predict with what probability a user will open a certain email. This model is used in a greedy way after certain prior logistic and heuristic delivery constraints are satisfied (Spark MLlib random forest regressor)
- *Topic Recommendations:* Treated sequences of webpage id's in a browsing session as words in a sentence and made page recommendations based on a Word2Vec model that was trained on the id sequences (gensim word2vec)
- *Hub-pages:* Created algorithm to organize community posts into categories like "troubleshooting," "device configuration," "product comparison," or undiscovered headings based solely on the post content (t-SNE, latent semantic analysis, Python nltk, affinity propagation, bag-of-word2vec, semi-supervised clustering)
- *Lookalike:* Developed a tool for expanding a group of users based on similar traits such that the "lookalike" users would be likely to take actions similar to those in the core group (association rules, frequent-pattern-growth)

Rice University — *Houston, TX*

Graduate Student — *Sept. 2009 - May 2015*

- Studied the geometry of spaces used for classifying geometric data, specifically point configurations
- Applied Hilbert-Mumford numerical criterion, Luna Slice theorem, Bialynicki-Birula decomposition
- Leveraged convex combinatorics to understand geometry of GIT semi-stable orbits

Research

ICERM: Control and Analysis of Large-Scale Time-Varying Data

Providence, RI

Researcher

August 6 - August 10, 2018

- Received funding to Collaborate@ICERM again this Summer, (PI: Deanna Needell, UCLA)
- Leveraged prior information to improve results for several compressed sensing recovery algorithms
- Explored recovery performance of multiple measurement vector algorithms when applied to time-varying signal models
- Experimented with adapting various multiple measurement vector algorithms to an online setting

ICERM: WiSDM Research Collaboration Workshop

Providence, RI

Researcher

July 17 - July 21, 2017

- Worked on a team of seven to produce results in compressed sensing and signal recovery
- Developed algorithms, simulations, and visualizations in python
- Provided theoretical guarantees for our algorithms

Papers

Compressed Sensing Research Collaboration

2017-Present

- N. Durgin, R. Grotheer, C. Huang, S. Li, A. Ma, D. Needell, and J. Qin. Compressed anomaly detection with multiple mixed observations. *Research in Data Science, Proc. WiSDM (ICERM)*, 2018
- N. Durgin, R. Grotheer, C. Huang, S. Li, A. Ma, D. Needell, and J. Qin. Randomized kaczmarz for support recovery of jointly sparse corrupted multiple measurement vectors. *Research in Data Science, Proc. WiSDM (ICERM)*, 2018
- J. Qin, S. Li, D. Needell, A. Ma, R. Grotheer, C. Huang, and N. Durgin. Stochastic greedy algorithms for multiple measurement vectors. 2017. Submitted

Service

IMA Math-to-Industry Bootcamp

Minneapolis, MN

Industry Mentor

July 20 - July 29, 2016

- Led a team of mathematics graduate students in the Grupo Bimbo Inventory Demand Kaggle competition
- Introduced supervised machine learning classifiers and relevant scikit-learn tools
- Advised on ways to improve training speed and model performance

BIG Math Network

Steering Committee

Sept. 2015 - PRESENT

- Seeking to strengthen connections between mathematical scientists in academia and those in business, industry, and government
- Providing both a graduate student and industry perspective on issues moving between the two spheres

Presentations

SIAM Annual Meeting AWM Workshop

Spokane, WA

Compressed Anomaly Detection With Multiple Mixed Observations

February 25 - March 1, 2019

Austin Big Data AI Meetup

Austin, TX

NLP Approach to Grouping Similar Questions in Community Forums

June 13, 2018