

Roni Kobrosly Ph.D.

DATA SCIENTIST · MACHINE LEARNING ENGINEER

📞 1-512-965-0601 | ✉️ roni.kobrosly@gmail.com | 🏠 www.kobrosly.net | 📺 ronikobrosly | 📷 ronikobrosly

Summary

An Austin-based, Machine Learning Engineer at Dia & Co. A former environmental health researcher with a decade of experience in machine learning and statistics. An enthusiastic, voracious learner with a hacker mindset. Excited about all things data, whether it applies to tech, politics, public opinion, or sports.

Work Experience

Charles Schwab

Austin, Texas

SENIOR DATA SCIENTIST

May 2016 - Sept 2018

- Identified previously undetected malicious and bot web client web sessions using MLlib's Doc2Vec and k-means clustering. Created big data pipeline to analyze 10TB batches of customer web log data using Spark.
- Prototyped novel cross-channel marketing attribution method using LSTM deep neural network with TensorFlow. Deployed on Google Cloud.
- Replaced \$150k/year Ipsos social media brand health reporting with an in-house tool created in Python. Obtained data through Twitter and Reddit APIs and custom-made news scraping pipeline. Employed topic modeling, sentiment analysis, and anomaly detection algorithms.
- Completed other high-impact projects with HR, customer support, and client experience teams. Employed causal inference modeling, regression analysis, and Bayesian inference.

Insight Data Science

New York, New York

DATA SCIENTIST AND PROGRAM DIRECTOR

Sept 2015 - May 2016

- Provided technical guidance and machine learning expertise to 66 unique Insight Fellow projects that included time series, supervised and unsupervised learning, and recommendation system components.
- Performed analyses that identified key factors that predicted Fellow success in program.
- Developed a Flask, D3, and Python-based internal dashboard to assess Fellow status and session KPIs.

Insight Data Science

New York, New York

FELLOW

June 2015 - Sept 2015

- Created SciClarify, a tool to help social scientists improve their chances of publishing in a high impact journal.
- Conducted natural language processing to derive features related to structure and syntax. Classified texts using logistic regression, support vector machines, and random forest. Obtained data through PubMed API, stored data in MySQL, processed with Python pandas.
- Built front-end with Flask, Bootstrap, and D3 and hosted web app on Amazon AWS.

Pure Earth / The Blacksmith Institute

New York, New York

RESEARCH SCIENTIST

Jan 2013 - Sept 2014

- Conducted the first-ever meta-analysis of Mexican environmental studies to estimate the true extent of lead exposure in children. Found that average blood lead levels were five times those of children in the US.
- Assisted staff with regression analysis and hypothesis testing for projects. Designed spatial survey methods to estimate the total number of industrial waste sites in Ghana.

Icahn School of Medicine at Mount Sinai

New York, New York

POSTDOCTORAL FELLOW

Sept 2012 - May 2015

- Applied machine learning techniques (multivariate adaptive regression splines, multiple regression, and random forest) to identify harmful chemical exposure among children, resulting in a total of 15 publications in top peer-reviewed journals.
- Created MySQL database to organize extensive laboratory data from 200,000+ patients from a large CDC study, which was used by multiple research groups at Mount Sinai, resulting in several publications.
- Predicted 3rd grade academic proficiency among 75,000 school children, with 78% cross-validated accuracy, in cooperation with the New York City Department of Health and Mental Hygiene.

Skills

Languages: Python, R, SQL, HiveQL, SAS, Stata, HTML, Javascript, Scala.

Tools: Jupyter Lab/Notebook, PyMC3, pandas, Dask, NumPy, SciPy, scikit-learn, Gensim, Flask, Matplotlib, Keras, TensorFlow, Vowpal Wabbit, Hadoop, Spark, MLlib, Docker, Rstudio, ggplot2, caret, git, UNIX, D3, Amazon Web Services, Google Cloud, PCF, Heroku, Neo4j, BayesiaLab.

Techniques: NLP (Word2Vec, Doc2Vec, GloVe), Bayesian Inference and simulation (MCMC, Monte Carlo), Optimization (Bayesian Optimization, Simulated Annealing, Genetic, Adam), Causal Modeling (Analysis of confounding, TMLE, SEM, Bayesian Networks), Sequence/time series analysis (LSTM, GRU, Attentional models, GAMS), Regression, Classification, Clustering (K-means, PAM), Reduction (SVD, PCA, Factor analysis, t-SNE).

Education

Ph.D. in Epidemiology, University of Rochester, Sept 2012

M.P.H. in Epidemiology, University of Michigan, May 2008

B.A. in Biological Sciences, Rice University, Dec 2005