Derek B. Yen

400 West 45th Street, New York, NY | derekyen@nyu.edu | +1 858-699-9684
Website: dbyen.github.io | LinkedIn: linkedin.com/in/derek-yen

Data Scientist with collaborative mindset, experience in analyzing large-scale data and building machine learning models in Python for entertainment/media, strong interest in data storytelling

---

Education

New York University — New York, NY
Master of Science in Data Science — May 2020

- GPA: 3.96/4.00
- *Coursework:* Deep Learning, Natural Language Processing, Machine Learning, Responsible Data Science, Probability and Statistics, Computational Linear Algebra and Optimization, Probabilistic Time Series Analysis

University of California, Los Angeles — Los Angeles, CA
Bachelor of Science in Mathematics of Computation, Minor in Neuroscience — June 2018

- *Coursework*: Information and Power, Algorithms in Bioinformatics, Mathematical Modeling

Skills and Qualifications

- **Programming**: Python and R (PyTorch, Scikit-Learn, NumPy, SciPy, Pandas, Tensorflow, RStudio, ggplot2)
- **Analysis & modeling**: Machine learning (Regression, Random Forest, Support Vector Machines, K-means), Deep learning (MLP, RNN, LSTM, CNN, BERT), Statistical analysis (Significance/hypothesis testing)
- **Big data & software**: SQL, Git/Github, Google BigQuery, Spark/PySpark, Databricks, Hadoop MapReduce
- **Visualization & reporting**: Tableau, Jupyter Notebooks, Matplotlib, Seaborn, ggplot2, LaTeX, Markdown, Microsoft Excel, Microsoft PowerPoint

Work and Research Experience

NBCUniversal Media — New York, NY
Data Scientist Intern — January 2019 – December 2019

- *Statistical testing and analysis*: Conducted statistical testing to measure differences in age group distributions over time, using PySpark for scalable data transformation in AWS Databricks environment. Designed visualization tool to simulate distribution changes using Retool and JavaScript. Fine-tuned linear regression models in R for forecasting demographic ratios
- *Evaluating forecasting models*: Collaborated with data engineering team to develop processes using MLFlow and R to save 800+ forecast models, evaluate based on MAPE, visualization, etc., and update MySQL tables. Designed app for senior data scientists to track model performance (folded into production process)
- *Developing forecasting models*: Produced 18-month forecasts with Python machine learning (ARIMA, regression, random forest) for shortform digital content including YouTube, Hulu, etc.

NYU Langone Health — New York, NY
Graduate Research Assistant — September 2019 – December 2019

- Applied multilabel text classification strategies to dataset of 6M+ NYU Langone Health medical notes with deep NLP models in PyTorch (BERT and XLNet) and GPU (CUDA) programming for capstone project

UCLA Mathematics Department — Los Angeles, CA
Undergraduate Research Assistant — September 2017 – August 2018

- Extracted and cleaned data from public census databases and geocoding APIs using Python
- Analyzed and visualized Los Angeles city regions with principal component analysis, nonnegative matrix factorization, correlation analysis, clustering, etc. studying homelessness data in Los Angeles
- Implemented models in Tensorflow for predicting large changes in homeless populations
- Communicated results to UCLA faculty and students through multiple presentations and reports

GuidedChoice — San Diego, CA
Customer Data Analytics Intern — September 2017 – November 2017

- Produced insightful visualizations with Python Matplotlib for Florida account to inform marketing strategy for retirement planning financial services firm

Swartz Center for Computational Neuroscience (University of California, San Diego) — San Diego, CA
Research Assistant — June 2015 – December 2017

- Wrote Python scripts to collect data on subject responses for experiments on neural correlates of competition
- Developed Matlab code for preprocessing EEG data using EEGLAB for K-means clustering, PCA, ICA
- Supported grant applications with detailed plots and statistical trend information

Selected Projects

- *News Article Text Classification*: Compared sentiment analysis methods (Naive Bayes, bag-of-words support vector machines, and neural networks) for set of articles from The New York Times
- *Recommender System with PySpark*: Used ALS for implicit feedback collaborative filtering in PySpark MLLib on Last.fm dataset. Experimented on cold starting using a latent feature regression model for unknown items
- *Building a Semantic Parser to Handle Queries about Song Data*: Applied the SEMPRE framework in Java to an interactive parser which can answer queries about a subset of the Million Song Dataset [Link]
- *Personal Podcast Analysis*: Using Python BeautifulSoup to webscrape and analyze RSS feeds from one year of my podcast listening history
- *Numerical Linear Algebra*: Implemented SoftImpute-ALS matrix completion algorithm in Python for recommendations with testing on MovieLens-100K

Leadership Experience

The Daily Bruin Los Angeles, CA

Copy Chief June 2016 – June 2017

- Managed scheduling, payroll, and development of the Daily Bruin style guide
- Responsible for ensuring content is edited for writing style, accuracy, and sensitivity on tight deadline for print (7K copies/day) and web (400K views/month) at award-winning UCLA student newspaper
- Led weekly meetings with seven slot editors and monthly training sessions with 30 contributors