# Fred Mubang, PhD

fredmubang@gmail.com **Machine Learning Researcher / Data Scientist** Linkedin: Fred Mubang
Website: fmubang.github.io

**Summary:** PhD Machine Learning Researcher and Data Scientist with 6 years of work experience. I am currently a Senior Data Scientist at Experian. Before that, I spent most of that time working for the Defense Advanced Research Projects Agency (DARPA), an agency within the Department of Defense. Also during that time, I was working on my Ph.D. in Computer Science from the University of South Florida, which I completed in October 2022. I specialize in Machine Learning algorithms such as XGBoost and neural networks, as well as NLP algorithms such as Transformers. I also specialize in data and statistical analysis.

## Work Experience

**Senior Data Scientist at Experian** **Oct 2022 - Present**

- Helped build a Large Language Model (LLM) for internal employee use. It's a chat bot that can aid thousands of Experian employees with understanding Experian's different datasets. Built using Transformer neural networks.
- Helped improve a data-driven algorithm using Pyspark and Python for finding the best contact information (phones and emails) for 200MM individuals for collections purposes. Achieved 30% lift for phone prediction and 20% lift for email prediction.
- Created a data-driven algorithm for calculating household size for healthcare Federal Poverty Line Prediction that incorporates Linear Regression and SMOTE for data augmentation. Achieved 40% lift. Saved organization $20+ million in potential lost revenue.
- Heavily utilized Python, Pyspark, Hadoop, Pandas, Scikit, and Numpy for data analysis and model building.
- Organized project timelines, set objectives, and led fellow team members through completion of project tasks
- Presented and communicated insights to upper management and stakeholders via Power Point presentations and reports.

**Machine Learning Researcher/Data Scientist for DoD/DARPA Social Simulation Project - Link** **Oct 2017 — Oct 2022 (5 yrs.)**

- Built machine learning algorithms (XGBoost, neural networks, LSTMs) to predict user activity for 50 million social media users on Twitter, YouTube, and Reddit. The goal was for DARPA to use these models for predicting the spread of misinformation online.
- Trained BERT (Transformer) neural networks for topic labelling tasks on various social media posts.
- Achieved 20% lift on historical baselines set by DARPA for these predictions.
- Performed various data engineering tasks such as cleaning, manipulating, scraping, feature engineering, and visualization of data.
- Performed detailed social network time series analysis of various datasets and created weekly Powerpoint presentations containing data analysis and insights. Used various Python libraries to prepare results such as Scikit, Pandas, Matplotlib, and Numpy.

## Education

**Ph.D in Computer Science,** University of South Florida, GPA: 3.87/4.00 **Aug 2018 — Oct 2022**

**Master of Science, Computer Science,** University of South Florida, GPA:3.87/4.00 **Aug 2018 — May 2021**

**Post Bachelor Studies, Computer Science,** University of South Florida, GPA:3.7/4.00 **Aug 2017 — July 2018**
**Bachelor of Arts, Music Business,** Berklee College of Music **Aug 2010 — May 2014**

**Relevant Courses:** Data Mining, Machine Learning, Neural Networks, Advanced Neural Networks, Social Media Mining, Network Science, Natural Language Processing, Intro to AI, Calculus 1-3, Linear Algebra, Probability and Statistics

## Skills and Technologies

- **General ML Skills:** Building ML models from scratch, Deep Learning, Neural Networks, Large Language Models (LLMs), ChatGPT, BERT, Transformers, Recurrent Neural Networks (RNNs), Long-Short Term Memory Networks (LSTMs), Generative Adversarial Networks (GANs), Convolutional Neural Networks (CNNs), Time Series Forecasting, XGBoost, Gradient Boosting, Probabilistic Modelling, Classification, Regression, Clustering, NLP, Word2Vec, Node2Vec, Feature engineering, Dimension reduction techniques, Logistic Regression, Linear Regression, Data Augmentation, SMOTE
- **Data Analytics:** Descriptive Statistics, Cleaning, Manipulation, Scraping, Visualization, Financial Modeling, Data Analysis
- **Business Skills:** Communicating Analysis, Leading group discussions, Mentoring, , Presenting Results Clearly to Stakeholders, Leading Business and Technical projects, Developing Analytical Solutions
- **Technologies:** Python, C, C++, Linux, Scikit-learn, Pandas, SQL, Pyspark, Hadoop, Tensorflow, Keras, XGBoost, Numpy, Seaborn, Matplotlib, Networkx, Jupyter Notebooks, Microsoft, Microsoft Powerpoint, Excel, Microsoft Office

## Publications (With Links)

- **Mubang, F.**, *Social Media Time Series Forecasting and User-Level Activity Prediction with Gradient Boosting, Deep Learning, and Data Augmentation.* Ph.D dissertation, College of Computer Science and Engineering, University of South Florida, USA, 2022. - Link
- **Mubang, F.**, Hall, L.O. *VAM: An End-to-End Simulator for Time Series Regression and Temporal Link Prediction in Social Media Networks. IEEE Transactions on Social Computing* (2022 - In Press) - Link
- **Mubang, F.**, Hall, L.O. *Simulating User-Level Twitter Activity with XGBoost and Probabilistic Hybrid Models. arXiv preprint arXiv:2202.08964* (2022) - Link
- **Mubang, F.**, Hall, L.O. *A Survey of Recent Artificial-Intelligence Driven Frameworks for User-Level Activity Prediction in Github. Engineering Applications of Artificial Intelligence* (2022 - Under Review) - Link