KISHORI M. KONWAR

MACHINE LEARNING ENGINEER / DATA SCIENTIST

(857) 214 9882 • kishori82@gmail.com • github.com/kishori82

SUMMARY

Experienced (5+ years) Senior Machine Learning Engineer with a strong background in developing and deploying advanced machine learning models. Proven track record in leading data science projects, optimizing algorithms, and collaborating cross-functionally to drive business solutions. Expertise in statistical modeling, neural networks, and large-scale data analysis.

SKILL

ML and Statistical methods

- Logistic regression
- Linear regression (lasso, ridge)
- Support Vector Machines (SVM)
- Decision Trees, Random Forest
- Neural networks, CNN
- XGBoost, Gradient Boosting
- K-Means Clustering
- Hierarchical Clustering
- EM Algorithm
- Dimensionality reduction, SVD, ICA, tSNE, UMAP, PCA
- Gaussian Mixture Model (GMM)
- Hidden Markov Models (HMMs)
- Statistical modeling

Technology/Platforms

- Hive, Presto, SQL, NoSQL
- Cloud computing, AWS, GCP
- ML model deployment

Machine Learning Frameworks

- Pandas, Scipy, ggplot, plotly
- PyTorch, scikit- learn, Tensorflow, Kera, CUDA

Programming skills

- Strong programming skills
- Linux, bash scripting, Perl, docker
- Python, R, data visualization
- C/C++, multithreaded programming
- Pipeline design & development
- Algorithms & data-structures
- Production software development

PROFESSIONAL EXPERIENCE

Data Scientist & Machine Learning Engineer| GateHouse Bio, Natick, MA — **May 2023 to April 2024**

- Computational and data science R&D, software development work for drug discovery. Created a deep-learning based method that detects potential small-RNA targets for neurological diseases.

Senior Research Data Scientist | Meta (formerly Facebook), Boston, MA — **Jan 2022 to Jan 2023**

- Developed machine learning (ML) algorithms to analyze large datasets and discover bottlenecks in user device response time. Provided solutions that led to 2x speed gain across 7 million user devices.
- Developed data pipelines and performed statistical analysis of large volumes of data to discover performance bottlenecks.
- Collaborated cross-functionally in a highly matrixed, virtual environment with global teams and agency partners on KPIs.

Data Scientist | Broad Institute of MIT & Harvard, Cambridge, MA — **Sep 2018 to Jan 2022**

- Developed statistical and ML techniques for analyzing single-cell data.
- Validated and developed new quantitative methodologies for comparison of various single-cell data.
- Designed, implemented and improved a large genomic data processing pipeline. The cost of processing samples (NGS) came down to $10,000 for processing 50,000 10x Genomics samples; previous methods cost $200,000..
- Analyzed large amounts of gene expression data (500 GB+) using a wide range of statistical and machine learning methods.
- Designed and implemented cloud-based pipelines for the Human Cell Atlas Project (single-cell transcriptomics) for large datasets.

Visiting Scientist | Research Laboratory of Electronics (RLE), MIT, Cambridge, MA — **Dec 2018 to Jan 2020**

- Developed algorithms for storage systems using erasure-codes.
- Developed error correction techniques for 5G networks.

Researcher | Computer Science and AI Lab (CSAIL), MIT, Cambridge, MA — **May 2015 to Sep 2018**

- Developed algorithms for distributed key-value storage systems using erasure-codes.
- Implemented an erasure-code-based distributed key-value storage system with strong consistency guarantees.

Post-doctoral Researcher | University of British Columbia, BC, Canada — **Jul 2010 to May 2015**

- Developed statistical methods for metagenomic data analysis.
- Developed pipelines for processing large next-generation sequence (NGS) metagenomics datasets.
- Developed data visualization software for large metagenomic datasets

Financial Engineer | KPMG, New York City, NY — **Oct 2008 to July 2010**

- Worked on mathematical modeling of various financial derivatives.
- Consulted large financial institutions on derivative valuations.
- Implemented mortgage valuation models at Freddie Mac 60x times faster than previous implementation.

Quantitative Developer | Goldman Sachs, New York City, NY — **Jan 2007 to Oct 2008**

- Validated and implemented risk models for credit, interest rate, equity derivatives.

- Implemented risk models for multi-billion dollar portfolios at the firmwide level.

SOFTWARE DEVELOPED

- **Optimus** • ***single-cell transcriptomics/genomics tool for NGS data*** • ***C++, multithreaded, cloud-based pipeline***
  Open-source, cloud-optimized pipeline developed at **Broad Institute** for the **Human Cell Atlas (HCA) Project** and **BRAIN Initiative Cell Census Network (BICCN)**. It supports the processing of any 3' single-cell and single-nucleus count data generated with the 10x Genomics. https://broadinstitute.github.io/warp/docs/Pipelines/Optimus_Pipeline/README/

- **MetaPathways** • ***Metabolic pathway and taxonomy predictor from NGS metagenomic data.*** • ***Python pipeline*** •
  Pipeline for predicting metabolic pathways and taxonomic groups in hundreds of metagenomic NGS sequences from ocean, land, human microbiome, and metagenomics, including niche environments. Currently, this tool is adopted by several Canadian research organizations. https://github.com/hallamlab/metapathways2/

- **MetaPathwaysGUI** • ***large data (500 GB+) visualization software for NGS metagenomics data*** • ***C++, Qt, multi-threaded visualization software*** • It can load over **500+** processed samples (each of 1GM data) with MetaPathways to interactively slice and dice with high responsiveness. This data-visualization tool is written in C++ using the Qt C++ platform. Runs on Linux, Windows and Mac. https://github.com/hallamlab/MetaPathwaysGUI

- **FAST** • ***aligner for genomic and transcriptomic sequences against large reference databases*** • ***C++ multithreaded*** •
  IO and multi-core optimized software performs DNA and amino acid homology search. FAST is over 2x the fastest available aligner and with 10x lower memory footprint. This enabled it to process several thousand samples on a desktop, which was considered impossible before. https://github.com/hallamlab/FAST

- **FragGeneScan-Plus** • ***gene finder on short NGS sequence data*** • ***C++ multi-threaded*** • This is an improved multi-threaded implementation of the **FragGeneScan** gene prediction model efficient in-memory data management to utilize multiple CPU cores without blocking I/O operations. This is 20x speed-up than the previous implementations.
  https://github.com/hallamlab/FragGeneScanPlusA

EDUCATION

Doctor of Philosophy (PhD) in Computer Science • **University of Connecticut, Storrs, CT**
Master of Science (MS) in Statistics • **University of Connecticut, Storrs, CT**
Master of Technology (MTech) in Computer Science • **Indian Statistical Institute, Calcutta, India**
Master of Science (MSc) in Physics • **Indian Institute of Technology, Kanpur, India**
Bachelor of Science (BS) in Physics • **Dibrugarh University, Assam, India**

RECENT PUBLICATIONS

(1) **A multimodal cell census and atlas of the mammalian primary motor cortex**, *Nature, Vol 598, 7 October 2021*

(2) **Composition and associations of the infant gut fungal microbiota with environmental factors and childhood allergic outcomes,** *mBio,* Volume 12 Issue 3, 2021.

(3) **Predicting metabolic modules in incomplete bacterial genomes with MetaPathPredict.** eLife 2024.

(4) **Pathway-centric analysis of microbial metabolic potential and expression along nutrient and energy gradients in the western Atlantic Ocean.** *Frontiers in Marine Science. 563, 2022*.

SELECTED DATA SCIENCE WORKS

(1) **ML predictive models for asthma development.** Created an ML method **(logistic-regression, SMOTE, imputation, etc.)** to co-associate with states of the bacterial composition of a baby and a rationally selected set of early-life environmental factors to its possible development linked to inflammatory diseases such as asthma at the age of five, with 81% accuracy. **(Journal: *mBio,* Volume 12 Issue 3, 2021)** https://journals.asm.org/doi/full/10.1128/mBio.03396-20

(2) **Explain Yourself: Why You Get the Recommendations You Do.** Rethinking the linear algebra in the current "spark.ml" **Collaborative Filtering (CF)** implementation, we demonstrated a way to explain the reason for the specific recommendation via the viewing history of a customer. We show how this is done and demonstrate its implementation as a new **"spark.ml"**, expanding the API. https://www.databricks.com/speaker/kishori-konwar. Databricks AI Summit presentation: https://youtu.be/BPzFFwYY0BM

(3) **Fuzzy decision tree, linguistic rules and fuzzy knowledge-based Neural Network:** generation and evaluation. Developed a **fuzzy knowledge-based neural network** based on the linguistic rules extracted from a **fuzzy decision tree**. Formulated a scheme for automatic linguistic discretization of continuous attributes. A new metric for the goodness of a decision tree in terms of its compactness (size) and efficiency is introduced. Effectiveness of the system was demonstrated on real-life data. Link

(4) **A Stationary Markov Chain Model for Labor Dynamics.** Labor market surveys usually partition individuals into three states: employed, unemployed, and out of the labor force. We model those labor paths as consecutive observations from independent Markov chains, where transition matrices are related to covariates through a multivariate logistic link. ***(appeared in Journal of data science. Volume 7, pages 27-42, 2008.)*** *Link*