Wayne Wang, PhD

email: wwang0328@gmail.com cell: 734.277.5499 web: ywa136.github.io

Summary

Data scientist and applied researcher with 6+ years of experience in statistics and machine learning. A wide spectrum of expertise in experimental design, causal inference, high-dimensional/spatio-temporal data analysis, optimization, and Bayesian inference. Proven ability to apply analytical methods to solve practical product problems with measurable impact on user engagement and revenue growth, develop innovative data-driven solutions at scale, and influence a variety of audiences (e.g., Product Managers, Engineers, Researchers, etc) in cross-functional projects.

Technical Skills

- **General Programming**: R, Julia, Python, MATLAB
- **Data Engineering**: SQL, BigQuery, Spark/Hadoop, Dataflow
- **Data Visualization**: R (tidyverse, dplyr, ggplot2, shiny), Python (bowtie, seaborn, plotly), Looker
- **Machine Learning Framework**: TensorFlow/TFX, scikit-learn, VertexAI
- **Cloud & Distributed Tool**: GCP, MPI, Git, Mercurial.

Professional Experience

- **Google**, Data Scientist *Jul '22 - present*
  - **Running and Analyzing Large-scale Online Experiments**: Designing A/B tests for YouTube Ads relevance measurement; defining user-centric metrics using online surveys and utilizing causal inference methods to drive statistically significant effects of ads relevance improvement; implementing scalable solutions to reduce bias in online surveys; optimally mapping users' sentiments to numerical scores via Bayesian optimization. *Tools: R, GoogleSQL* (☞Constructed metrics focused on ads relevance and successfully landed as secondary `launch metrics` for YouTube Ads overall marketplace optimization).
  - **Developing Model-based Measurement Framework**: Developing and productionizing machine learning models (e.g., DNNs, Decision Forests) for measuring and optimizing users' perceived relevance towards ads. *Tools: R, Python, GoogleSQL* (☞Increased metric sensitivity in A/B experiments by `2x to 12x`, effectively saved the team from having to increase the survey load by at least `4x`, which amounts to roughly `53 million` ads slot saved at minimum; Developed associated internal `R packages` for general model-based metric construction).
  - **Designing and Analyzing Human Evals**: Designing human evaluation templates for contextual ads relevance deep learning model training and evaluation. Implementing statistical models (e.g., mixed-effects models, Krippendorff's alpha) for measuring and improving label/eval quality. *Tools: R, GoogleSQL* (☞Revamped evaluation templates and measurement techniques have boosted the agreement among human raters by more than `3X`, `effectively reducing` evaluation cost and the noise of the ratings subsequently used in machine learning model training).
- **Los Alamos National Laboratory**, Research Intern *May '21 - Aug '21*
  - **Streaming Distributed PCA for Exascale Climate and Space Sciences**: Designed an communication-efficient streaming & distributed PCA algorithm for online analysis and visualization of exascale data generated from climate and space weather simulations. *Tools: Julia, MapReduce, MPI* (☞Paper published at ACM/IEEE Supercomputing Conference '21; Developed an associated open-source Julia package called `TributaryPCA`).

Research Experience

6+ years of research experience with 10+ peer-reviewed publications in top statistical journals and machine learning conferences with 60 citations (☞Google Scholar page). Selected research projects include:

- **High-dimensional Gaussian Graphical Models for Tensor-Variate Data**: Proposed a novel statistical model for high-dimensional multiway/tensor-variate data. Designed efficient optimization algorithms for learning the underlying parameters. Improved downstream task such as ensemble Kalman filtering and image classification. *Tools: Julia* (☞Papers published at AISTATS '20, ICML '21, NeurIPS '21, and Statistics Surveys; Developed an open-source Julia package called `TensorGraphicalModels`).
- **Time-Varying Topic Models**: Developed a framework for topic modeling of time-varying corpora, combining parametric statistical models with nonparametric computational geometric methods. *Tools: Python, Spark, Hadoop* (☞Paper published in Harvard Data Science Review; Developed an online exploratory analysis/visualization tool using `R Shiny`).
- **Bayesian Point Process Models**: Developed a novel point process model for tracking the onset of extreme events (e.g., earthquakes, solar flares) and designed an efficient Bayesian inference methods for parameter estimation. *Tools: R, Stan.*
- **Deep Learning for Solar Flare Forecasting**: Proposed an ensmeble method combining LSTM and CNN for classification of flare-imminent active regions using video data. *Tools: Python* (☞Paper in The Astrophysical Journal).

Education

- **University of Michigan** Ann Arbor, MI
  Ph.D. in Statistics *Sep '18 - Jul '22*
  Dissertation: Interpretable and Scalable Graphical Models for Complex Spatio-temporal Processes
- **University of British Columbia** Vancouver, Canada
  M.S. in Statistics *Sep '16 - Aug '18*
- **Simon Fraser University** Vancouver, Canada
  B.S. with Distinction in Actuarial Science (Completed SOA Exams P, FM, and MFE) *Sep '12 - May '16*