DONG ZHANG, Ph.D.

Seattle, WA
(614) 397-5173 | dongzhanghz@gmail.com
github.com/dongzhang84
linkedin.com/in/dongzhanghz

- 15+ years of research and industry experience in machine learning and AI modeling, computer vision, natural language processing, cloud computing, parallel/distributed computing and numerical simulations. Excellent skills of critical thinking, creative problem solving, project management, team building, and written/oral communication.
- 16 peer-reviewed publications (with 14 first-author) on major journals, total citations 1000+. Invited referee for four top international journals.

PROFESSIONAL EXPERIENCE

- **Applied Scientist, Amazon — Seattle, WA** February 2022 –
  - Work within the Buyer Risk Prevention (BRP) group, developing **tree-based** and **neural network** models to detect and prevent buyer fraud across multiple businesses world widely.
  - Developed and maintained 20+ ML models in production to prevent various types of fraud across Amazon business from online retail and **Just-Walk-Out** physical stores.
  - Created a global unified model and developed on Sagemaker that significantly reduced the non-payment fraud rate for the Try Before You Buy business across the US, multiple EU countries, and Japan by 50%.
  - Led a research project using a **multi-task learning** to consolidate different types of fraud detection into a single neural network model to boost model performance. Results summarized in a paper and published in Amazon internal ML workshop.
  - Fine tuned **Large Language Models (LLMs)** to predict classification tasks which outperform baseline tree-based model, and applied fine-tuned LLM in fraud detection.
  - Led a **MLOPs** project of building a dashboard for automatic monitoring and refresh of models in production, collaborating with software engineers to significantly reduce the operational burden on scientists for model maintenance.
- **Data Scientist, IBM — Durham, NC** March 2020 – February 2022
  - Led project to build a **regression** model to predict onsite work order duration for entire IBM technical support onsite repair services. Developed MLOps to automatically update and retrain the model.
  - Conducted a **text classification** model for Kenexa BrassRing email system using Linear SVC (baseline), cloud computing (Watson NLC), and other ANN models.
  - Built a new NPS Early Warning System detractor **binary classification** model for the IBM support ticketing system, significantly improved recall by 70% compared to the previous model. Analyzed the model impact using statistical hypothesis testing.
  - Improved a **content-based recommendation** system for IBM business partners by re-building the data engineering pipeline and implementing time series prediction as a new feature.
  - Collaborated with IBM AI Research team for the Symbolic AI Deep Learning project.
  - Mentored 6+ junior data scientists and other IBMers for their career path and data science training.
- **Data Science Fellow, Insight Data Science — Seattle, WA** September 2019 – December 2019
  - Created a tool *Classify3D* for a tech startup to automatically segment and identify objects from 3D point-cloud images. Used **unsupervised ML** (DBSCAN and GMM) to cluster images, and **computer vision** feature detector (ORB) to compare image similarities. Identified several classes of objects above **95%** accuracy. Developed the tool as a web app using **Flask** and **Docker**.
- **Research Scientist, University of Michigan — Ann Arbor, MI** September 2018 – August 2019
  - Developed **high-performance computing** simulations using half million CPU-hours to study multiple astrophysical processes. Generated ∼1TB 3D synthetic data for image processing and hypothesis testing.
- **Research Associate, University of Virginia — Charlottesville, VA** September 2015 – August 2018
  - Led two **parallel computing** radiation hydrodynamic simulation projects in C/C++ using ∼2 million CPU-hours on various supercomputers. Generated ∼10 TB synthetic data for statistical modeling.
  - Developed **computer vision** tool and created pipeline in Python to visualize ∼1 TB multidimensional data generated from simulations. Analyzed data using linear/polynomial regression, correlation and classification.
  - **Optimized algorithms** to perform the most accurate simulations for astrophysical radiative systems, which can be observed by multi-wavelength ground and space telescopes.
- **Research Assistant, The Ohio State University — Columbus, OH** September 2009 – July 2015
  - Built physical models using (semi)-analytic methods to explain up-to-date observations of hundreds of galaxies.
  - Led independent projects to develop new models of dark matter structure to explain the origin of early Universe.

SKILLS

I am proficient in end-to-end ML engineering with the following skills:

- **Programming Languages**: Python, SQL, Spark, C/C++, R
- **Machine Learning**:
  - ◇ All core skills in **Supervised Learning** & **Unsupervised Learning**. Tools: Scikit-learn, SageMaker, MLlib.
  - ◇ **Deep Learning** including ANN, CNN, RNN, LSTM, GAN, Transformer, Diffusion Models, Transfer Learning, Few-Shot/Zero-Shot Learning. Tools: PyTorch, Tensorflow, openCV, relevant Github/HuggingFace packages.
  - ◇ **Natural Language Processing** & **Large Language Models**. Tools: NLTK, Gensim, spaCy, Transformers, HuggingFace LLM foundation models, openAI GPT.
  - ◇ **Time Series Forecasting**. Tools: statsmodels, Fbprophet, self-developed NN models.
  - ◇ **MLOPs**. Tools: Weights & Bias, MLflow, Docker, Kubernetes.
- **Statistics**: Hypothesis Testing, A/B Testing, Estimation, Bayesian Inference, Simulations
- **Database**: MySQL, Postgres, SQL Server, Db2 Big SQL, Cloudant NoSQL DB
- **Cloud Computing**: Amazon Web Services (SageMaker), Google Cloud Platform, IBM Watson Cloud.
- **Miscellaneous**: Data Analysis (pandas, SQL, spark), Visualization (seaborn, matplotlib, plotly, Tableau, generative AI), Product Deployment (Flask, Docker, Cloud Service), Web Scraping.

SELECTED SIDE PROJECTS

- ***Generative AI Comic Books and Films***: Use state-of-the-art text-to-image and image-to-video generators along with large language models (LLMs) to create stories and transform these stories into comic books and films. You can view the selected book and film projects at this link: snowboatstudio.com.
- ***On the Temperature of ML Systems***: Developed a thermodynamic framework for machine learning (ML) systems, integrating physical concepts such as temperature, energy, and entropy to interpret model training and refreshment as a process of state phase transition. Neural networks can be viewed as complex heat engines with variable temperatures across layers, fundamentally governed by the principles of thermodynamics. (arxiv.org/abs/2404.13218).
- ***New Business Model for Generative AI Tools***: Proposed a prompt-based scoring system as a complete new revenue-sharing business model between AI tools such as ChatGPT and their data providers. Sharing revenue between AI tools and their data providers could transform the current hostile zero-sum game relationship between AI tools and a majority of copyrighted data owners into a collaborative and mutually beneficial one. (arxiv.org/abs/2305.02555).
- ***Data Entropy Analyzer***: original research on data entropy using statistical mechanics and information theory to determine structure of multi-dimensional data, developed a tool to calculate data entropy and monitor data shifting (github.com/dongzhang84/Entropy_Analyzer).
- ***Book recommender system***: created content-based book recommender system with users' review data scrapped from Goodreads, using NLP word embeddings and cosine similarity comparison. Deployed model on AWS using Flask (github.com/dongzhang84/BookReco).
- ***Gas turbulence driver***: wrote C++ code to generate 3D turbulence in gaseous medium by Fast Fourier Transform. Analyzed turbulence data using Gaussian distribution and correlation, and visualized turbulence evolution (e.g., youtu.be/-W0y6wHAU2w, youtu.be/O1u1Jgd2148).

EDUCATION

- ◇ **Ph.D.,** astrophysics (GPA 4.0/4.0) — Ohio State University, Columbus, OH, *July, 2015*
- ◇ **M.S.,** astrophysics (GPA 3.8/4.0) — Nanjing University, Nanjing, China, *June, 2009*
- ◇ **B.S.,** astronomy, *Summa Cum Laude* (GPA 1/45) — Nanjing University, Nanjing, China, *June, 2006*