DONG ZHANG, Ph.D.

US permanent resident | Durham, NC   github.com/dongzhang84
(614) 397-5173 | dongzhanghz@gmail.com   linkedin.com/in/dongzhanghz

10+ years of research and industry experience in statistical/machine learning modeling, predictive modeling, programing, data visualization, computer vision, numerical simulations, cloud computing, parallel/distributed computing, critical thinking, problem solving, project management and excellent communications.

SKILLS

- **Programming Languages**: Python, C/C++, SQL, R, bash, MATLAB, HTML
- **Machine Learning/Statistical Modeling**:
  - Supervised Learning (linear/logistic regression, regularization, SVM, naive bayes, k-NN, decision tree, random forest, XGboost, LightGBM, cross validation)
  - Unsupervised Learning (k-means, GMM, DBSCAN, hierarchical clustering, PCA)
  - Deep Learning (CNN, RNN, LSTM), Computer Vision
  - Natural Language Processing (word embedding, TF-IDF, SVD, LSA, topic clustering)
  - Hypothesis Testing, A/B Testing, Time Series Analysis/Forecasting
- **Tools**: Scikit-learn, Tensorflow, Keras, Pandas, NLTK, Gensim, Numpy, Scipy, Matplotlib, Seaborn, Fbprophet, Jupyter, PySpark, BeautifulSoup, Flask, openCV
- **Cloud Computing**: Amazon Web Services (EC2, S3), IBM Cloud (Watson Studio, AutoAI, Watson Assistant), Google Cloud, Docker
- **Parallel/Distribution System**: MPI, OpenMP, Spark (RDD, spark.sql, spark.ml)

EXPERIENCE

- ***Data Scientist***   March 2020 – present
  IBM — Durham, NC
  - Analyzed the impact of a detractor recommendation model of the IBM support ticketing system on customer support team using **statistical hypothesis testing**.
  - Built a new NPS Early Warning System detractor prediction baseline model using **classification algorithms** XGBoost and LightGBM to significantly improve the accuracy from old model recall $\sim$0.36 to recall $\sim$0.6.
  - Led a project to develop a **supervised regression** model to predict work order duration of the IBM technology support services. Aggregated data from IBM Cognitive Enterprise Data Platform (CEDP) to improve model accuracy. Collaborated with software engineers to implement the model as a **production**.
  - Mentored new data scientists/analysts and other IBMer for their career paths and data science training.

- ***Data Science Fellow***   September 2019 – December 2019
  Insight Data Science — Seattle, WA
  - Consulted for a tech company, created a tool **Classify3D** to automatically segment and identify objects in 3D images. Each image is **2-10 GB** with **50-100 million** data points (dongzhang84.github.io/classify3d).
  - Used **unsupervised ML** tools (DBSCAN and GMM) to cluster 3D images, and used **computer vision** tool openCV to compare image similarities. Identified several classes of objects above **95%** accuracy. Developed the frontend web app using **Flask** and HTML.

- ***Research Scientist***, Computational Astrophysicist   September 2018 – August 2019
  University of Michigan — Ann Arbor, MI
  - Developed **high-performance computing** simulations using **half million CPU-hours** to study multiple astrophysical processes in supernova remnants. Generated $\sim$**1 TB** 3D HDF/VTK data from simulations.

- ***Research Associate***, Computational Astrophysicist and Data Engineer
  University of Virginia — Charlottesville, VA   September 2015 – August 2018
  - Led two **parallel computing** radiation hydrodynamic simulation projects written in C/C++ using $\sim$ **2 million CPU-hours** on various supercomputers. Generated $\sim$**10 TB data** for image processing.
  - Developed **computer vision** tool and created pipeline in Python to visualize $\sim$**1 TB** multidimensional data generated from simulations. Analyzed data using linear/polynomial regression, correlation and classification.
  - **Optimized algorithms** to solve radiative transfer equations to perform the most accurate simulations for many radiation systems, which can be observed by multi-wavelength ground and space telescopes.

- ***Research Assistant***, Theoretical Astrophysicist   September 2009 – July 2015
  The Ohio State University — Columbus, OH
  - Built synthetic 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.

SELECTED SIDE PROJECTS

- **Book Recommender System**: Created content-based book recommender system using users review data scrapped from Goodreads, by NLP word embeddings (**TF-IDF**, **SVD**, **LSA**, **doc2vec**), and cosine similarity comparison. Flask used to deployed the model on AWS (github.com/dongzhang84/BookReco).
- **Kaggle Competition: Predict Future Sales**: Built Lasso and XGBoost **supervised machine learning regression models** with cross validation to predict future sales of a software firm.
- **Flight Delay Predictor**: Built a tool to predict flight delays for various airports/airlines, using **supervised classification** models including logistic regression, Gaussian process, random forest, kNN, and SVM. (github.com/dongzhang84/Flight_delay)
- **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.,** in Astrophysics (GPA 4.0/4.0)	Ohio State University, Columbus, OH, *July, 2015*
◇ **M.S.,** Astrophysics (GPA 3.8/40)	Nanjing University, Nanjing, China, *June, 2009*
◇ **B.S.,** Astronomy, *Summa Cum Laude* (GPA 1/45)	Nanjing University, Nanjing, China, *June, 2006*

SYNERGISTIC ACADEMIC ACTIVITIES

- ○ 15 peer-reviewed publications on major journals with 14 first-author publications, total citations 650+.
- ○ Reviewed papers as a referee for four top international journals.
- ○ Presented 30+ invited talks at colloquia and seminars, and contributed talks over 10+ international conferences.