CHUNLIN HE, PhD
Data Scientist

chhe09@gmail.com • (765) 490-7331 • West Lafayette, IN • [LinkedIn](#) • [Github](#) • [Website](#)

SUMMARY

Results-oriented Data Scientist with a Ph.D. in Plant Science specializing in Biostatistics, an MSc in Biostatistics and a Data Science certificate from Microsoft. Proficient in Python, R, SQL, KQL, SAS, with expertise in Machine Learning and Deep Learning. Basic understanding of AWS services through training. Skilled in optimizing predictive models, conducting thorough assessments, and refining models for improved business performance. Experienced in data wrangling, statistical analysis, and various tools for data science. Adept leader in managing end-to-end data projects, collaborating on health industry initiatives, clinic data analysis, and implementing relational database management systems. Accomplished in statistical and quantitative research projects, with a notable portfolio of 58 peer-reviewed papers and book chapters. A dynamic problem-solver poised to excel in a Data Scientist role.

EDUCATION

- Professional Certificate: "Large Language Models: Foundation Models from the Ground Up", Databricks, USA.
- Data Science Certificate: Microsoft Professional Program (MPP) for Data Science, Microsoft, USA
- Professional Certificate: Data Science Fundamentals, Microsoft, USA.
- Ph.D.: Plant Science with a focus on Biostatistics, University of Saskatchewan, Canada.
- M.Sc.: Biostatistics and Quantitative Genetics, Nanjing Agricultural University, Nanjing, China.
- B.Sc.: Crop Science, Hunan Agricultural University, Changsha, China.

SKILLS

- Programming Languages: Python, R, SQL, KQL, SAS, JavaScript.
- Machine Learning (ML) and Deep Learning (DL): Linear Regression, Logistic Regression, Random Forest, Decision Tree, SVM, Gradient Boosting (AdaBoost, XGBoost, CatBoost), KNN, Neural Networks, NLP & LLM.
- Data wrangling and statistical analysis: EDA, Hypothesis Testing, Correlation and Regression Analysis, Chi-square Test, ANOVA, t-test, Statistical Modeling, Principal Component Analysis, Nonparametric Statistics.
- Tools and Technologies: Github, AzureML, Power BI, Spark, Tableau, HDInsight, AWS, MS SQL Server, MySQL, SQLite, RStudio, Kusto.Explorer, Jupyter Notebook, VS Code, PyCharm.

PROFESSIONAL EXPERIENCE

Freelance Data Scientist, West Lafayette, IN   **Aug 2023 - present**

- Extract, clean, and preprocess large datasets for analysis.
- Analyze and interpret complex business data to identify trends, patterns, and opportunities.
- Conduct statistical analysis and hypothesis testing to support decision-making processes.
- Develop and implement machine learning models for predictions.

Modeller/Data Scientist, Tata Consultancy Services Ltd., West Lafayette, IN   **Sep 2020 – Jul 2023**

- Conducted comprehensive assessments of predictive models, identifying gaps for improved model risk management and business performance.
- Specialized in developing and refining models related to business operations, with a focus on predicting customer satisfaction and membership retention.
- Evaluated model accuracy, weaknesses, and limitations through rigorous testing and benchmarking, resulting in significant improvement in predictive performance.
- Implemented effective model maintenance, monitoring, and documentation practices, contributing to continuous business performance improvements.

Microsoft/MPP Data Science Track & Freelance Data Scientist, W. Lafayette, IN   **Jan 2019 – Aug 2020**

- **Health Industry Collaboration (NDA):** Conducted data mining and database management for a private health company, ensuring data integrity and accuracy. Implemented data cleansing, query operations, and merging processes for large data sets.
- **Clinic Data Analysis (NDA):** Analyzed extensive datasets from a private research organization using R, offering valuable insights derived from statistical analysis. Data sets aggregated from tens of thousands of participants supported decision-making and enhanced organizational outcomes.
- **Relational Database Management (RDBMS):** Implemented a successful project for a private company, establishing a relational database management system (RDBMS) using MS SQL Server. Executed data manipulation, storage, and retrieval operations.
- **Microsoft/MPP Machine Learning Projects:** Demonstrated expertise through the successful completion of diverse machine learning projects including classification and regression. Addressed real-world data science challenges commonly faced by corporations and research organizations.

Alfalfa Toolbox/Data Curator, Noble Research Institute, Ardmore, OK   **May 2015 – Dec 2018**

- **Project Leadership:** Led data curation and analysis for the Alfalfa Toolbox project, contributing to the acceleration of molecular breeding for the 8-billion-dollar alfalfa market.
- **Data Pipeline and Analysis:** Executed end-to-end data pipeline and exploratory data analysis (EDA) using Python and R on diverse datasets from multi-environmental trials, extracting actionable insights.

- **Database Management:** Designed and maintained databases with RDBMS (MS SQL Server), specializing in efficient extraction of alfalfa populations with multiple traits for the Alfalfa Toolbox project, partners, and customers.
- **Integration Optimization:** Enhanced data integration into the Alfalfa Toolbox web portal by optimizing processes through the Toolbox API, ensuring seamless access and usability.

Breeding Services Manager – Genomic Data/Senior Scientist, GCP/CIMMYT, Texcoco, Mexico

Oct 2010 – Dec 2014

- Spearheaded and coordinated over 100 data projects, overseeing the generation of genomic and genotypic data for 100+ academic and research institutions spanning 50+ developing countries.
- Applied advanced data analysis techniques, such as QTL mapping, to extract actionable insights that facilitated the integration of genotypes and phenotypes for international breeding programs.
- Mentored and guided a team of research scientists, providing expertise in utilizing cutting-edge data analytics tools and creating visualizations for the effective analysis of genotypic and phenotypic data.

REFERENCES

Available upon request.