CHUNLIN HE, PhD
Data Scientist, Microsoft Certified

West Lafayette, IN 47906 • 765-490-8041 • chhe09@gmail.com • https://datute.net

SUMMARY

- Managed databases, curated and analyzed data for the multi-million dollar alfalfa toolbox project (ABT) to accelerate molecular breeding targeting an 8-billion dollar alfalfa market.
- Managed 100+ data projects in generating genotypic data for the 100+ academic and research institutions across 50+ developing countries.
- Strong experience and knowledge in data science projects with computing skills in Python, R, SQL, Bash, SAS, Minitab, MS SQL Server, MySQL, NoSQL (MongoDB), RDBMS, Jupyter Notebook, Google Colab, RStudio, AzureML, Power BI, Tableau, HTML5, CSS, and Big Data analytics using MS R Client, HDInsight, Spark.
- Experience and knowledge in data wrangling, Machine Learning (ML) and Deep Learning (DL) including Linear Regression, Logistic Regression, Random Forest, Decision Tree, Support Vector Machine, Gradient Boosting, AdaBoost, XGBoost, CatBoost, Neural Networks, computer vision, image analysis.
- Authored and co-authored 58 peer-reviewed and conference papers including book chapters.
- Other knowledge and interest: GBS, RNA-seq analysis, molecular genetics, genomics and bioinformatics.
- Citizenship: United States.

EDUCATION

Data Science Certificate: Microsoft Professional Program (MPP) for Data Science, Microsoft, USA.
(Certificate No: b5b3db2d-82dc-4aec-a623-37d86279b0b3, awarded in 2020, no expiration date)

Professional Certificate: Data Science Fundamentals, Microsoft, USA.
(Certificate No: 0731b691e08a473e97dc95a49633df27, awarded in 2019, no expiration date)

Ph.D.: Plant Science with Biostatistics, University of Saskatchewan, Saskatoon, Canada.

M.Sc.: Biostatistics and Quantitative Genetics, Nanjing Agricultural University, Nanjing, China.

B.Sc.: Crop Science, Hunan Agricultural University, Changsha, China.

SKILLS AND KNOWLEDGE

- **Computing skills:** Python, R, MS R client, SAS, Minitab, SQL, Bash/Unix, JavaScript, AzureML, Power BI, Spark, Tableau, HDInsight, MS SQL Server, MySQL, MongoDB, RDBMS, HTML, CSS, Google Colab, Jupyter Notebook.
- **Machine Learning (ML) and Deep Learning (DL):** Linear Regression, Logistic Regression, Random Forest, Decision Tree, Support Vector Machine, Gradient Boosting, AdaBoost, XGBoost, CatBoost, Neural networks, computer vision, image analysis.

- **Data munging and statistical analysis:** Hypothesis testing, correlation and regression analysis, Chi-square test, ANOVA, t-test, statistical modeling, principal component analysis, nonparametric statistics.
- **Bioinformatics:** Genome sequence assembly, annotation, DNA sequence analysis, gene identification, RNA-seq, gene expression analysis, sequence motif, SNP markers, DNASTAR, GeneMapper, MapQTL, QTL IciMapping, Biopython.

SOCIAL/MEETUP GROUPS: (1) SQLSaturday; (2) IndyPy; (3) IndyUseR Group; (4) Indy Azure User Group; (5) Power BI User Group, (6) Indy Big Data, (7) Data Science Indy; (8) Big Data Developers in Indianapolis.

PROFESSIONAL EXPERIENCE

Freelance Data Scientist 2019 - present

Recent Data Science Projects

- **Kaggle deep learning project**: global detection of wheat spikes/heads using computer vision and image analysis.
- **Data merging and database management:** data cleansing, query and merging, exploring their relationships and database management with a private company through NDA in health industry.
- **Kaggle machine learning project:** data analysis, visualization and prediction of the coronavirus (Covid-19) pandemic, aiming to take measures to contain the pandemic.
- **Clinic data analysis:** accomplished successfully a statistical analysis using R for a set of clinic data collected from several thousand participants from a private research organization and provided valuable insights for decision-making.
- **Relational database management project:** carried out successfully a RDBMS project from a private company using MS SQL server.
- **Kaggle machine learning project:** classifying the forest cover types using Python with an accuracy of 87.85% - 100% using the algorithms of logistic regression, Support Vector Machine (SVM), Random Forest, XGBClassifier.

Microsoft Professional Program (MPP) for Data Science Track 2018 - 2019

- **MPP machine learning projects:** (1) predicting using regression model on customer budget spending, (2) classifying item purchases with an accuracy of 85% - 100%.
- **Microsoft Professional Capstone Challenge:** successfully accomplished an MPP Capstone Challenge using data set with 44 features hosted by DrivenData to accurately predict the national-wide gross rents in USA. Algorithms applied include (1) Linear Regression, (2) Random Forest, (3) AdaBoost, (4) Decision Tree.
- **MPP Capstone Challenge Project II:** successfully predicted the mortgage rate spread across 50 states from a training data set with a number of independent variables using the following machine learning algorithms: (1) Linear Regression, (2) Random Forest, and (3) XGBoost.

Alfalfa Toolbox/Data Curator, Noble Research Institute, Ardmore, OK 2015 - 2018.12

- Curated and analyzed alfalfa data for the multi-million dollar toolbox project (ABT) to accelerate molecular breeding targeting the 8-billion dollar alfalfa market.

- Generated insights through data wrangling, Exploratory Data Analysis (EDA) for the diverse data sets from multi-environmental trials using Python and R.
- Designed and updated databases using RDBMS (MS SQL Server) and extracted essential populations of alfalfa with multiple traits of interest from the databases using SQL for ABT, partners and customers.
- Optimized the integration of data into the ABT web portal through the Toolbox API.

Breeding Services Manager/Senior Scientist, GCP, c/o CIMMYT, Texcoco, Mexico 2010 - 2014

- Managed 100+ data projects in generating genotypic data for the 100+ academic and research institutions across 50+ developing countries.
- Created numerous insights from genotypes to phenotypes through data analysis including QTL mapping for many international breeding programs.
- Mentored many research scientists through demonstrating data analytics tools and visualization of genotypic and phenotypic data.

Corn Breeder/Senior Scientist, Dow Chemical/Dow AgroSciences, Fowler, IN 2007 - 2009

- Successfully generated and collected diverse, multi-environmental high-quality data sets for the corn breeding program for yield performance and disease resistance.
- Performed data extraction, manipulation and statistical analysis using in-house tools over large relational data sets from various field trials.
- Strategically made data-driven decisions and successfully selected and advanced many superior inbreds and hybrids with high yield and quality.

Corn Breeder/Scientist, Monsanto Company, Gothenburg, NE 2005 - 2007

- Successfully designed experiments and generated many data sets including agronomic and disease data of corn breeding populations.
- Statistically analyzed all data from field trials using existing analytical platform and created insights for extracting outperforming inbreds and hybrids.
- Made data-driven decisions, successfully selected and advanced many pre-commercial hybrids.

Research Geneticist/Postdoc Associate, USDA-ARS, SGIL, Beltsville, MD 2002 - 2004

- Developed SSR and SNP markers and generated genotypic data for alfalfa germplasm.
- Successfully accomplished genotypic data analysis using SAS and generated distinctive dendrogram.
- Made important genomic data comparisons regarding the germplasm relationships.

Postdoctoral Fellow, Agriculture and Agri-Food Canada (AAFC), Harrow, ON, Canada 2000 - 2002

- Developed SSR markers and generated genotypic data for tomato disease resistance.
- Successfully accomplished data analysis using SAS, identified and mapped the disease resistant genes.

- Made valuable data-driven decisions and selected high yielding lines with multiple disease resistance.

AWARDS/HONORS

- Recognition of Contribution Award/Plaque, GCP/CIMMYT, El Batan, Mexico 2014
- Performance Award, GCP General Research Meeting, Lisbon, Portugal 2013
- Outstanding Team Member Award for disease screening, Monsanto, Gothenburg, NE 2007
- Certificate of Merit Award, SGIL, USDA-ARS, Beltsville, MD 2003

RESEARCH GRANTS

- **He, C.** and X. Delannay. 2012. Addition of a forensic component (QA/QC) to the TL1 and TL2 programs. Bill & Melinda Gates Foundation, Texcoco, Mexico.
- **He, C.** and X. Delannay. 2011. High density genotyping for elite and popular varieties of GCP target crops, Bill & Melinda Gates Foundation, Texcoco, Mexico.
- Poysa, V., **C. He** and K. Yu. 2001. "Development of molecular markers to facilitate development of disease resistant processing tomatoes for Ontario", funded by Ontario Tomato Research Institute (OTRI) and the Matching Investment Initiative (MII) fund of Agriculture & Agri-Food Canada (AAFC), Harrow, Canada.

SELECTED MEETING PRESENTATIONS

- **He, C.,** N.N. Diop and G. McLaren. 2014. Development and Utilization of KASP SNPs for Molecular Breeding in the Developing Countries. International Plant and Animal Genome XXII Conference, January 11-15, San Diego, CA, USA.,
- **He, C.** 2013. Genotyping for the Forensics Project (QA/QC). Annual Tropical Legumes I (TLI) Project Meeting (TLM), 22-24 May, Kampala, Uganda.
- **He, C.** 2013. Fingerprinting Update and Visualization of Data/database. Annual Tropical Legumes I (TLI) Project Meeting (TLM), 22-24 May, Kampala, Uganda.
- **He, C.** 2012. Forensics Project. Annual Tropical Legumes I (TLI) Project Meeting (TLM), 7-11 May, Addis Ababa, Ethiopia.
- **He, C.** and X. Delannay. 2011. Genotyping Services for Molecular Breeding in Developing Countries: Opportunities and Perspectives. In Proceedings of ASA, CSSA, and SSSA Annual Meeting, Oct. 16-19, San Antonio, TX.

BOOK CHAPTER

- **Chunlin He,** John Holme, Jeffrey Anthony. 2014. SNP Genotyping: The KASP Assay. In Delphine Fleury & Ryan Whitford (eds.), *Crop Breeding: Methods in Molecular Biology*, Humana Press, pp. 75-86 (Chapter 7).

PEER-REVIEWED JOURNAL PUBLICATIONS

A list of 18 peer-reviewed papers (available upon request in PDF format).

CONFERENCE ABSTRACTS

A list of 39 conference papers (available upon request).