Tayler A. Blake

Columbus, OH
☎ *(740) 607-9508*
✉ *tayler.a.blake@gmail.com*
http://taylerablake.github.io

Education

May 2018 **Ph.D. Statistics**, *The Ohio State University*, Columbus, Ohio.
Advisor: Yoonkyung Lee

January 2010 **M.S. Statistics**, *The Ohio State University*, Columbus, Ohio.

May 2007 **B.A. Mathematics, Computer Science**, *Capital University*, Columbus, Ohio.

Experience

Director, Data Science and Analytics

OCLC, *Columbus, OH.*

August 2023-July 2023

Worked closely with product management and technology leaders to influence and align the business needs to technical strategy to develop the vision and execution of OCLC's approach to data science and analytics. Drove growth and maturation of OCLC's new internal data science practice, from identifying business opportunities to leveraging data science and machine learning to build predictive models and leading their implementation and deployment. Leads three teams comprising the Data Science and Analytics group having the collective mission to unlock the value contained in the company's extensive data ecosystem in service of our academic, public, and global library customers.

The Analytics team provides self-service data capabilities for internal OCLC stakeholders, developing enterprise data models, monitoring and measuring data quality in addition to creating and maintaining data reports and dashboards for both external customers as well as internal customers companywide. As director, served as technical lead for two major initiatives during my OCLC tenure. The first of which leveraged dbt alongside a number of SQL variants including Spark SQL, PostgreSQL and MySQL to develop an enterprise cloud (Snowflake) data warehouse in an effort to centralize the company's wide variety of disparate data assets residing in an amalgamation of storage technologies. The effort required a breadth of sophisticated ELT and ETL techniques to migrate unstructured document data, including internal operations data extracted from Jira and Confluence and OCLC's global network of bibliographic records. This project was executed in tandem with an enterprise-wide migration all of all company reporting dashboards from Oracle's Business Objects platform to Microsoft's PowerBI.

Served as technical lead for two of Data Science's major initiatives, the first of which entailed performing record linkage on a huge volume of bibliographic records. These data are the backbone of almost the entirety of OCLC's product portfolio, necessitating data quality improvement by removing as many redundant records (records which refer to the same library asset) as possible. The modeling effort required leveraging both Natural Language Processing (NLP) techniques (word and paragraph embeddings) as well as supervised learning techniques to deliver a scalable classification model that identifies pairs of duplicate records. The solution had to scale to accommodate the growing volume of OCLC's global network of bibliographic records, on the order of hundreds of millions. Additionally, oversaw the use of Large Language Models (LLMs) and Retrieval Augmented Generation (RAG) to automate or streamline the creation of XML stanzas for customer libraries employing an OCLC product that facilitates library patron authentication via SASL and TLS.

Founding Data Scientist, Privacy Engineer

Ketch, *San Francisco, CA.*

September 2021-March 2023

Served as engineering's privacy expert and lead research and development in statistical privacy and security. As one of the first members of the company's product and technology organization and the first data science hire, lead the development of Ketch's implementation of privacy enhancing technologies including k-anonymization and mechanisms guaranteeing differential privacy; developed several components of Ketch's core data governance product in Golang including its native integration with Snowflake, AWS Redshift, Oracle, MySQL, Postgres, and MariaDB database technologies. Spearheaded the data science practice within the engineering organization, and owned the data science effort backing the classification model underlying Ketch's software product for the discovery of sensitive data and PII (personally identifying data) which leveraged Python and Tensorflow.

Senior Research Scientist

Immuta, *Columbus, OH.*

September 2020-September 2021

Senior member of research team driving product feature development for data access and governance software. As Immuta's expert in re-identification risk quantification, lead research and product development of a comprehensive policy-agnostic disclosure risk and utility measurement framework for facilitation of automating privacy policy recommendations for maximum data utility for given disclosure risk tolerance levels. Using this research, prototyped Immuta's data asset risk scoring feature in Python and Javascript.

Contributed to Immuta's digital catalogue of technical and white of ebooks and white papers on topics in statistical disclosure control including methods in risk quantification and recovering basic statistical models with anonymized inputs. Lead efforts in designing and prototyping performance testing for the Immuta/Databricks native integration using Python, PySpark, and Spark SQL

Senior Data Scientist

Redjack, *Columbus, OH.*

September 2019-September 2020 Lead consulting data scientist supporting federal government cyber defense efforts. Extracted structured and unstructured data from relational databases (MySQL and PostgreSQL) and MongoDB to train supervised machine learning models to identify spam email campaigns. Leveraged model interpretation techniques implemented in Python to understand the best features for discriminating between phishing and non-phishing emails and how those inputs affect the probability of spam.

Data Scientist

Root Insurance, *Columbus, OH.*

April 2019-September 2019 Data scientist supporting teams responsible for customer acquisition and development of pricing plans. Designed and analyzed A/B tests for assessing the impact of marketing campaigns and pricing and underwriting changes on customer acquisition and retention. Extracted and transformed raw log data from a variety of cloud data storage sources including Amazon Redshift and RDS, Snowflake, and MariaDB to develop conversion and retention models for predicting the impact of pricing changes on KPIs including loss ratio and total bound premium. Integrated with product teams and actuaries at every stage of the data science life cycle, from problem formulation and data collection to model deployment and communication of insights generated from analysis using Tableau, Python, and RMarkdown.

Senior Consulting Data Scientist

Information Control Company, *Columbus, OH.*

January 2017-March 2019 As a consulting senior data scientist, lead data science team members including junior analysts, data engineers, and visualization specialists in planning, designing, and delivering data science solutions addressing a variety of business questions for clients across several industries including retail, food and beverage, insurance, and marketing across a variety of tech stacks, both on-premise and cloud-based. Project work required execution of the entire data science lifecycle, from gathering business requirements and documentation, ETL and data warehouse construction, SQL-based data aggregation and profiling, exploratory data analysis and model training and tuning in R and Python, data visualization and dashboard reporting in Tableau, PowerBI, R and Python, and model deployment and monitoring.

Internally at ICC, contributed to Advanced Analytics development seminars and coordinated the Advanced Analytics journal review, curating and leading discussions about current literature in statistics, data science, and machine learning and its application to ICC client problems.

Machine Learning Specialist

Pillar Technology, *Columbus, OH.*

February 2016-December 2016 Machine learning specialist on a team responsible for designing and developing adaptive software for an IoT product to be embedded in a line of luxury vehicles. Challenged to design and implement machine learning models in the absence of data a priori, leveraging regularization methods and programmatic model selection techniques. Owned data science efforts from initial exploratory analysis to production deployment of models in Java and Scala using Agile development best practices.

Spearheaded components of project planning pertaining to data and the corresponding infrastructure necessary for data collection, storage, and modeling at scale. Lead efforts to establish client trust in data modeling and algorithms, a new operating space for Pillar.

Data Scientist

Store Development, Starbucks Coffee Company, *Seattle, WA.*

May 2014 - November 2015

Data scientist on a team serving the company in market planning and strategy. Utilized a variety of statistical and machine learning methods, both supervised and unsupervised such as penalized regression and classification, generalized linear models, ensemble methods including boosting, bagging, and random forests, with applications of the latter to both classification and clustering with application to a product recommendation engine, predicting consumer total lifetime value, sales forecasting, and estimating cannibalization effect of newly launched locations on existing stores.

Using heterogeneous spatial and temporal data from disparate sources, estimated the causal impact of a variety of interventions, such as competitor store openings, pricing changes, and new product launches on store performance using Bayesian structural time series models and summarizes comparisons between this methodology and prior approaches, including the traditional difference-in-differences estimators. Created dynamic visualization dashboards in R Shiny to communicate insights gleaned from exploratory analysis, model summaries and performance.

Adjunct Statistics Instructor

Department of Mathematics, Columbus State Community College, *Columbus, Ohio.*

August 2013 - January 2014

Lecturer for an introductory statistics course for undergraduate students, Statistics 1350. Non-instruction responsibilities included curriculum and assessment development, including lecture presentations and online learning tools and learning assessments.

Graduate Research Assistant

Comprehensive Cancer Center, The Ohio State University, *Columbus, Ohio.*

October 2011 - August 2012

Responsibilities included analysis of large microarray data sets, in particular utilizing data mining and dimension reduction techniques to find genetic markers in leukemic patients, sharing results and collaborating with medical professionals to both direct further laboratory investigation as well as further statistical investigation.

Graduate Research Assistant

Nationwide Center for Advanced Customer Insights, The Ohio State University, Nationwide Insurance, *Columbus, Ohio.*

June 2010 - June 2011 Responsibilities included work on projects modeling agency behavior using high dimensional demographic and marketing data, specifically by modeling survival times using Cox proportional hazards models with both static and time-varying coefficients. Modeling was done with an emphasis on building parsimonious, interpretable models. I was responsible for presenting results in a corporate setting to high level company executives with motive to encourage and motivate business decisions and action.

Research Interests

My current research interests are centered in privacy enhancing technologies and statistical disclosure control. I am particularly interested in how we assess the quality of statistical disclosure control methods and how to quantify the risk of re-identification associated with data to which these methods have been applied. Quantifying this risk is challenging for a number of reasons, including the breadth of the mathematical mechanisms underlying these techniques and how to justify the assumptions about the resources available to any potential adversary. My most recent research proposes a framework for estimating the probability of re-identification that is agnostic to privacy model under the most conservative assumptions about an attacker's background knowledge and available resources.

Computing Skills

Machine Learning and Data Modeling:

R, Python (Jupyter & JupyterLab, Sci-kit Learn, Sci-Py, Pandas, Numpy, Matplotlib, Plotly), PySpark, SparkML, Snowflake, Databricks, MLFlow, Tableau, PowerBi, dbt

General Software Development:

Golang, Python, PySpark, SQL (Postgres, MySQL, Spark SQL), Unix shell, Flask, Docker, AWS (EC2, RDS, Redshift, Elastic Beanstalk, S3, CLI), Protocol Buffers (protobuf), basic knowledge of C++, Java and HTML

Other:

Git/version control, LaTeX, project management software including Jira and Confluence

Presentations

December 2022 **Toward a Unifying Information-Theoretic Framework for Re-identification Risk Quantification**, *IMS International Conference on Statistics and Data Science*.

August 2022 **Statistical Privacy: No Free Lunch**, *Superset Super Summit*.

August 2018 **Smoothing spline ANOVA models for nonparametric covariance estimation for longitudinal data**, *Joint Statistical Meetings*.

August 2017 **Nonparametric covariance estimation for longitudinal data via tensor product smoothing**, *Joint Statistical Meetings*.

October 2016 **The Machines Are Coming: Will Algorithms Replace Designers in the UX World?**, *Wards Auto User Experience Conference*.

August 2012 **Nonparametric Covariance Estimation for Functional Data with Shrinkage Toward Stationary Models**, *Joint Statistical Meetings*.

Honors and Awards

June 2010 Ohio State University Department of Statistics Teaching Assistant of the Year Nominee

June 2009 Ohio State University Department of Statistics Teaching Assistant of the Year

June 2008 Ohio State University Department of Statistics Teaching Assistant of the Year Nominee

References

Available upon Request