

Ethan Chen

ethanchen@hsph.harvard.edu
linkedin.com/in/ethan-chen-052804280

Education

Harvard University	Aug 2026 - Dec 2027 (Expected)
Master of Science in Health Data Science	

University of Illinois at Urbana-Champaign (UIUC)	May 2026
Bachelor of Science in Econometrics and Statistics	

- **GPA:** 3.98/4.00 Dean's List Honor (All terms)
- **Relevant Coursework:** Statistical Modelling (2 Semesters), Statistics and Probability (2 Semesters), Statistics Programming Methods, Financial Econometrics
- **Training:** Oeson Learning Internship: Completed a 12-week remote intensive training in real-world data analytics through 4 technical modules (Advanced Excel, Descriptive Statistics, Python, SQL), delivering 3 end-to-end projects: Global Burden of Disease Study (58,905 records, K-Means Clustering), Stock Market Analysis (1,279 data points, Regression & Volatility Modeling), and Film Investment Strategy (1,000+ movies, SQL Query Optimization)
- Served in the Econ Tutoring Center for 5 Months

Internship

- | | | |
|--|------------------------------------|---------------------------|
| The Insights Partnership, Chicago, United States | Marketing Research Intern | Feb 2026-Jun 2026 |
| <ul style="list-style-type: none">● Supported a mixed-method brand refresh project for Lincoln University by synthesizing 1:1 interviews, focus groups, and survey findings into client-ready reports and executive presentation materials.● Built and organized qualitative and quantitative research assets, including quote banks, verbatim appendices, profile slides, and mindset-based analyses, while validating key survey figures and improving reporting consistency.● Developed storyline-driven slides and early video/storyboard drafts using participant quotes, charts, and cross-segment insights, helping stakeholders review findings in a clearer and more actionable format. | | |
| GE Motors by Wolong, Indiana, United States | Data Analyst Intern | May 2025-Aug. 2025 |
| <ul style="list-style-type: none">● Developed a regression model in R analyzing OEM vs. Distributor unit price disparities across states, product divisions, and volume tiers, visualized trends via faceted ggplot2 boxplots, identifying 3 outlier orders underpriced by \$130K, presented findings to leadership, informing pricing strategy adjustments and the prevention of pricing errors● Led the digitalization initiative for internal sales request process, iterate form structure, decision logic, and approver role settings, designed and implemented a multi-level approval system using Microsoft Power Apps, the system significantly reduced approval ambiguity in day-to-day sales operations● Cleaned and restructured pricing Excel files containing 40+ technical fields (e.g., thrust ratings, voltages, model identifiers, coupling keys) to help generate more readable pricing catalogs, collaborated with engineering/marketing head to verify the changes | | |
| WeFIRE, Silicon Valley, United States | Product Data Analyst Intern | Dec 2024-May 2025 |
| <ul style="list-style-type: none">● Conducted user behavior analysis for the App by using Python (pandas, scikit-learn) to process 190000+ records, performed data cleaning, EDA, and feature engineering to build a Random Forest model that identified attribute A as the key conversion driver, linked to a 3x activity boost, the model achieved 74% accuracy and 84% recall, forming the basis for data-driven marketing optimization● Researched 6 competitor products, conducting in-depth competitive analysis through weekly detailed reports with a strong focus on AI functionalities—the hallmark of our product | | |
| ATLAS Program, Illinois, United States | Data Analyst Intern | May 2024-Dec 2024 |
| <ul style="list-style-type: none">● Built an interactive Tableau dashboard to track computer usage by class, presented findings to the UIUC Data Leadership Team to support strategic decision-making● Produced 18 charts for the DMI department, analyzing international student distribution which revealed key insights into demographic trends and campus diversity | | |

- Manipulated Jan. – Jun. 2024 sales data (6000+) using Quick BI and updated 25 multi-department dashboards
 - Conducted ad hoc reporting and analysis using SQL subqueries and joins to support sales channel expansion
- Initiated contract redundancy detection using SQL, the automated reminder reduce managers' time by 30%

Research

Research Assistant, School of Molecular & Cellular Biology, UIUC Professor Kevin Van Bortle Jan 2025-Jan 2026

- Designed and implemented multiple gene expression normalization methods (TPM, RPKM, CPM, Quantile, Min-Max) with S4 generic compatibility for SummarizedExperiment objects in R
- Followed Bioconductor standards to develop modular and reusable functions, ensuring code quality with no nested helpers and achieving 100% test coverage, created comprehensive examples and interactive RMarkdown vignettes demonstrating function usage with airway RNA-seq data
- Conducted regression testing to validate consistency between new and legacy results, and built workflow visualizations (ggplot2, flowcharts) for downstream interpretation
- Collaborated with lab members to prepare package documentation and compliance checks, supporting submission to the Bioconductor ecosystem

Research Assistant, Department of Statistics, UIUC Professor Jeffrey A Douglas Jan 2025-Jan 2026

- Analyzed a large-scale 16PF personality dataset with over 50,000 responses, experimented with advanced latent-variable methods such as Latent Class Analysis (LCA), Limited Information Factor Analysis (LIFA), and Multidimensional Item Response Theory (MIRT), conducted model selection using AIC, BIC, ICL, and CAIC, and evaluated model fit with indices including CFI, TLI, RMSEA, and SRMR
- Implemented iterative modeling workflows in R, automating large-scale simulations and convergence checks to optimize parameter estimation
- Interpreted factor structures and class memberships to compare competing personality models, bridging empirical findings with theoretical frameworks (16PF vs. Big Five)
- Prepared visual outputs including scree plots, eigenvalue distributions, and factor loadings heatmaps, making results interpretable for both technical and non-technical audiences

Research Assistant, Dona Francisca Health Study - Econ data lab, UIUC Professor Flavia Andrade Dec 2023-Feb 2025

- Played an instrumental role in the Relationship between Old Adults' Mental Health Outcome and Living Conditions
- Selected Longitudinal Aging Study in India (LASI) and Chinese Longitudinal Healthy Longevity Survey (CLHLS) as our datasets from Gateway to Global Aging database based on predefined parameters and target variable
- Implemented a logistic regression and addressed class imbalance issues in the LASI to gain more robust predictions
- Presented inconsistent findings across the datasets at the UIUC Undergraduate Research Symposium (Apr 2025)

Project

Core Member, Equity in post-HCT Survival Predictions (Kaggle Competition, Bronze Medal) Dec 2024-Mar 2025

- Developed a novel loss function incorporating algorithmic fairness constraints, ensuring balanced model performance across 6 different racial groups, reduced the standard deviation of prediction performance between groups from 0.025 to 0.018, significantly enhancing model fairness by 28% and addressing potential biases in healthcare AI
- Designed and executed a multi-level model ensemble strategy, integrating 13 heterogeneous models (including deep learning, XGBoost, LightGBM, etc.). Through weighted rank fusion, the final ensemble model achieved a C-index of 0.621 on the private test set while achieving a 66% improvement in inference speed

Skills

- Data Analysis & Modeling: Causal Inference, Regression Analysis, Hypothesis Testing,
- Machine Learning: Regression, LASSO, Time Series Analysis
- Programming: R(Advanced); Python, SQL (Intermediate); Microsoft Office Tools; Adobe Tools
- Data Visualization: Tableau (Graphs, Dashboards), R (ggplot, plotly, shiny)
- Spoken Languages: Chinese (Native), English (Fluent), Shanghainese (Fluent)