Andrew Wheeler, PhD

Raleigh, NC | LinkedIn | GitHub | Personal Website | apwheele@gmail.com

TECHNICAL SKILLS

Code: python (10+ years), R (10+ years), SQL (10+ years), ArcGIS, GitHub, Tableau, JavaScript, PHP
Statistical Modeling: Machine Learning (Random Forests, Boosted Models), Deep Learning, Mixed Integer Linear Programming, Network Algorithms, Spatial Analysis, Time Series forecasting, and Regression/Causal Inference
Generative AI Experience: I have written a book, *Large Language Models for Mortals*, as an introduction to using foundation model APIs across all major providers (OpenAI, Anthropic, Google, and AWS). Additionally, it has a chapter on LLM coding tools (GitHub Copilot, Claude Code, and Google Antigravity).
Python Libraries with Strong Experience: sklearn, scipy, pandas, sqlalchemy, pulp, geopandas, matplotlib, pytorch, huggingface, CatBoost, boto
Supervisory Experience: Supervise multiple junior data scientists and have developed in house classes for data science. Previously supervised masters' theses and PhD student projects.

For tech and modeling experience, I can point to external examples (peer-reviewed publications, blog posts, or open-source GitHub contributions) to demonstrate competence in **every claimed domain.**

EDUCATION

PhD, SUNY Albany, Criminal Justice (2008-2015)
Masters, SUNY Albany, Criminal Justice (2008-2012)
Bachelors, Bloomsburg University, Criminal Justice (2004-2008)

EXPERIENCE

Gainwell Technologies, Director, May 2025 to Current

Director of Data Science in Gainwell's Artificial Intelligence group. Example projects I have contributed to and oversee are in payment integrity, coordination of benefits, population health, and prior authorization. Example recent projects involve

- Main architect of smart index labelling for medical records using Anthropic models on Bedrock. Reduces nurse review times for DRG and POS Medicaid claims by 5~10 minutes (over 1 million dollars savings per year)
- Improving claim processing in COB direct bill product, improving revenue by an estimated $4 million in 2026
- Main architect of creating our master person index horizontally across applications, going from 3 weeks to under 1 day to run, and projected increase of $5 million dollars of revenue per year.
- Using a technique I developed to limit false positive rates with conformal sets to limit prior authorization auto-approvals to an error rate of less than 1% in Wisconsin.

Gainwell Technologies, Principal Data Scientist, December 2019 to May 2025

End to end data scientist, working on predictive models to identify problematic health insurance claims and develop models for population health monitoring. Example projects include:

- Developed a model to identify overpayments in health insurance claims. Deployed model generates over $8 million additional revenue per year.
- Main developer on population health metrics (readmission risk, non-urgent ER usage, future cost) for multiple state clients on Databricks
- Created an algorithm to help identify service desserts that is currently being implemented for the state of California

University of Texas at Dallas, Assistant Professor of Criminology, July 2016- December 2019

I have published over 30 peer reviewed articles over my career. Additionally, I have created courses related to data analytics and data visualization for both undergraduates and graduates. My personal blog highlights many of these technical applications, as well as illustrates my proficiency in writing and presentation. Several example projects include

- Identifying hot spots of crime that generate over $1 million dollars in labor costs for police.
- Helping Carrollton PD redraw their patrol beats to equalize officer workload and reduce travel to calls by 20%.
- Creating a fairness algorithm to help prioritize hotspots of crime while limiting racial inequality expected to occur from police contacts.

Finn Institute for Public Safety, Research Analyst, January 2010- July 2016

The Finn Institute conducted research in partnership with police departments in upstate New York. Example projects I collaborated on while an analyst at the Finn Institute were:

- creating a social network algorithm to prioritize offenders in gang interventions
- Automating the generation of individuals to call for post police contact surveys
- Eval. the predictive accuracy of a chronic offender tool (simple models vs machine learning).

SOFTWARE EXPERTISE AND EXAMPLE PROJECTS

Python (10+ years, focus on network statistics, linear programming, and machine learning)

- My book, *Data Science for Crime Analysis with Python*.
- Using pytorch to build latent class mixture models
- Synthetic control in python: Opioid death increases in Oregon and Washington, using custom Lasso implementation and conformal inference
- Predicting algae blooms via satellite data, 2nd place in DrivenData competition ($9,000 in winnings)
- Using linear programming to create optimal allocation with network spillovers

R (10+ years, focus on machine learning, data visualization, and spatial statistics)

- *ptools* package CRAN, GitHub
- Managing R environments using conda
- Age, period, cohort graphs for suicide and drug overdose using ggplot

SQL

- Binomial confidence intervals in SQL, and discrete time Kaplan-Meier estimates in pyspark SQL
- Creating a memoized function in Postgres and JavaScript
- Profiling RAM consumption in pandas reading in chunks and several databases

GitHub

- Caching huggingface models in GitHub actions
- Dallas crime dashboard and Raleigh crashes chart automated in GitHub actions run on a crontab schedule
- Setting up pyspark to run SQL tests in GitHub actions

Tableau

- Temporal analysis (Seasonal Charts, Weekly time series with error bars)
- Example Crime Analysis Dashboard

JavaScript and PHP

- Network prioritization tool runs all client side in JavaScript (due to sensitive data)
- Sworn dashboard is built using D3.js and Supabase backend
- PHP + google sheets backend for a custom survey, can use query encoding to route to custom surveys (S1, S2)

SELECTED DATA SCIENCE PUBLICATIONS

Jacques, S & **Wheeler, AP** (2025) A plea for open access to qualitative criminology: With a Python script for anonymizing data and illustrative analysis of error rates, *The Journal of Qualitative Criminal Justice and Criminology* 15(2): 634d1d80

- Using a huggingface model and fuzzy linkage, we identify sensitive information in qualitative narratives, and apply an algorithm to contextually replace names, areas, etc. with pseudonyms in the text.

Circo, G & **Wheeler, AP** (2023) Using Every Door Direct Mail Web Push Surveys and Multi-level modelling with Post Stratification to estimate Perceptions of Police at Small Geographies.

- This is the winning solution to the NIJ Challenge on Innovations in Measuring Community Attitudes for the non-probability sampling approach. We suggest sending a QR code on a mailer via every-door-direct-mail, and then use multi-level regression with post-stratification to correct for differential response bias and small samples. This approach is very cost effective compared to in person canvassers, and is more geographically targeted than online approaches. Total winnings of $25,000.

Circo, G & **Wheeler, AP** (2021) National Institute of Justice Recidivism Forecasting Challenge Team "MCHawks" Performance Analysis.

- This technical report provides a description of our submission to the *NIJ Recidivism Forecasting Challenge*, in which our predictive solution placed on the leader board for 7 different categories and we collected just under $40,000 in prizes. We discuss our solution to meeting the racial fairness constraints, and suggest alternative metrics for future competitions that are likely to be less volatile to optimize.

Wheeler, AP & S Reuter (2021) Redrawing hot spots of crime in Dallas, Texas. *Police Quarterly* 24(2): 159-184.

- I use an unsupervised clustering technique (DBSCAN) to identify cost of responding to crime hot spots. I find compared to the current hot spot areas implemented by Dallas PD, my identified areas are much smaller, and capture cost of crime at much higher densities. One hot spot I identified has over a million dollars of crime cost per year, suggesting a hot spot policing strategy is likely to have appreciable return on investment for Dallas PD.

Wheeler, AP & W Steenbeek (2021) Mapping the risk terrain for crime using machine learning. *Journal of Quantitative Criminology* 37(2): 445-480.

- I spatially forecast long term robbery hot spots in Dallas using random forests and several other common techniques. I find random forests are much more accurate than other state of the art (e.g. RTM). I also use interpretable machine learning summaries to evaluate several different criminological theories and provide local summaries for individual hot spots.

Wheeler, AP (2020) Allocating police resources while limiting racial inequality. *Justice Quarterly* 37(5): 842-868.

- I tackle the problem of how hots spots policing exacerbates disproportionate minority contact, and construct a linear program intended to balance police targeting of hot spots, while constraining the number of minorities likely to be stopped by the police.

Wheeler, AP, SJ McLean, KJ Becker, & RE Worden (2019) Choosing representatives to deliver the message in a group violence intervention. *Justice Evaluation Journal* 2(2): 93-117.

- I create a greedy social network algorithm to identify individuals who should be targeted for a gang intervention, which the motivation is to spread the deterrence message to the remaining gang members. I use simulations to show it often finds the minimal dominant set for networks the typical size and density of gang networks.

Wheeler, AP, RE Worden, & JR Silver (2019) The predictive accuracy of the violent offender identification directive (VOID) tool. *Criminal Justice and Behavior* 46(5): 770-788.

- I evaluate the predictive accuracy of a scoring system created to forecast violent gun offenders. I compare the accuracy of the ad-hoc tool developed by the police department, relative to logistic regression models and machine learning models.

Wheeler, AP (2019) Creating optimal patrol areas using the P-median model. *Policing: An International Journal* 42(3): 318-333.

- I formulate an integer linear program, with constraints on workload equality, to re-draw patrol beats for the Carrollton, TX police department. My results find my beats are likely to be over 20% more efficient in reducing drive time to calls for service compared to the current beat layout.

SELECTED CLASSES TAUGHT

Crime Mapping (Graduate Level), online course materials here

- Using ArcGIS, GeoDa, and R, I teach principles of geographic analysis, geographic data visualization, and spatial econometrics using examples from crime analysis

Seminar in Research & Design (PhD Level), online course materials here

- common quasi-experimental research designs (propensity score matching, fixed/random effects, differences-in-differences, synthetic-control), and other analysis (e.g. missing data imputation, mixture models, social network analysis, machine learning)
- Provide code examples in R, Stata, SPSS, and python for each week

Crime Analysis (Undergraduate), online course materials here

- Time series monitoring and forecasting, and geographic mapping techniques in Excel
- Advanced Pivot Tables, Interactive Graphics, and Dashboard creation in Excel
- SQL queries and relational databases (Access)