Boris Dev

boris.dev@gmail.com

Interests

- Domain modeling: translating domain expertise into code
- Process optimization: debugging data hand-off toil, human-in-the-loop design, uncovering extraneous cognitive load, tracking down wrong assumptions that grow in the cracks that split teams.
- Building observability systems: monitoring, KPIs, and analytics
- Learning: Ethereum and causal inference

I am looking to pull a company's stuck data project out of the quicksand. I want to fix a messy problem that requires working with engineers, subject matter experts, and data scientists. I want to save you $$$. My toolkit includes cognitive empathy, agile code refactoring, building observability systems, and inviting feedback with demos and papers (examples).

Coding tools

- Product: Linux, Python, Postgres, Flask, Django
- DevOps: Docker Compose, Kubectl, Kafka, Celery, Helm, AWS Sagemaker, AWS Lambda
- Data science and analytics: SQLAlchemy, Pandas, Numpy, PyTorch, Tableau, AWS Groundtruth
- Instrumentation and observability: Splunk, LightStep, ELK, Grafana, Prometheus

Job experience

Data Scientist / Technical AI Product Manager consultant at SimpleLegal, 2022 - 2023

The company's first AI feature was stuck. As tech lead, my puzzle was to figure out why the performance of their machine learning models for text classification was flat even after the company had been spending more money on human annotation. The actions I took resulted in the company launching their first AI feature, with positive feedback from customers and the sales team. Below are the sequence of actions that I took.

- I stopped the human annotation process (pulled the Andon Cord).
- I identified an incorrect assumption: our performance blocker was not training data quantity, but quality. Then I identified two main culprits of the poor training data quality: 1) convoluted annotation guidelines and 2) missing pre-processing noise filters.

- I immersed myself in the metrics to perform triage on eleven machine learning classifier models: one was never needed; two were replaced by expert rules; five were sufficient; and two were targeted for relabeling.
- I immersed myself in the company's legal invoice data through exploratory data analysis and by labeling several thousand sentences while continuously getting feedback from our subject matter expert (SME).
- I simplified the annotation guidelines in collaboration with both the SME and annotation team.
- I designed a new annotation Human-in-the-loop ML QA process in collaboration with the SME and annotation team. This included a CI (continuous improvement) process where the annotators, our SME, and myself reached consensus to fix the guidelines as we hit edge cases (ie. feedback).
- I added pre-processing noise filters to the labeling pipeline (AWS GroundTruth).
- I wrote papers to explain new concepts and changes for the product team and executives.
- I worked daily with our NLP-ML expert on re-prioritization of R&D work (ie. triage).
- I added a new QC process (embarrassment review sheets and staging server).
- I refactored the inference server (AWS Sagemaker) with new post-processing, decoupling, thresholds and preprocessing noise filters.
- I assigned Jira issues to the engineers and data scientist.

Backend developer at Sight Machine, 2018 - 2021

- brought the company's biggest public facing feature at the moment, Recipes from its embryonic start as a spreadsheet to general release.
- non-technical explanations to the product and customer support teams
- analytic endpoints (Flask, Numpy, Pandas, Celery, SqlAlchemy)
- high level design papers (feature, Policy Based Access Control)
- started the company's first distributed tracing using LightStep.
- simplified our development environment.
- distributed system debugging using Kubectl, Helm Charts, and Grafana.

Data and product engineer tech lead at HiQ Labs, 2015 - 2018

- owned web scraping to get around LinkedIn's bot detection system
- ran and tracked experiments on different spider configurations.
- led developers and data scientists to move from a monolithic pipeline to a microservice pipeline (Spark, Kafka, Rancher, Docker Compose).
- refactored old code so it could be decoupled into separate services, achieving horizontal scaling, reduced tech debt and reduced cognitive load during releases.
- designed and built the company's first observability system
- trained data scientists on Kafka and Spark, and junior developers on

coding.

Start-up partner and developer at Map Decisions, 2014

I created a mobile app to automate street sign inspection (Angular, Django)

Developer at Urban Mapping, 2011 - 2013

- built the company's first observability system (Splunk and Tableau)
- built the company's first Jenkins QA CI system

Open source and work papers

- Co-founder of library for clustering geographic areas, github.com/clusterpy.
- A play Ethereum MEV bot, github.com/borisdev/play_mev_bot
- A git bare approach to version control your dot files, github.com/borisdev/dotfiles
- Work papers

Academics

- Ph.D dissertation: Assessing Inequality using Geographic Income Distributions 2014.
- Entry in Encyclopedia of Human Geography on Spatial Econometrics. Sage Publications. 2009
- Interactive spatiotemporal modelling of health systems: the SEKS–GUI framework
- sigma-convergence in the presence of spatial effects
- Integrating Econometric and Input-Output Models in a Multiregional Context