

Jack Stehn

Machine Learning Engineer | Data Engineer | Data Scientist

San Francisco, California (US)

✉ jack@ste.hn

LinkedIn: jack-stehn GitHub: jstehn ste.hn



SUMMARY

High-agency Data Professional with a unique background blending rigorous social science research with production-grade engineering. My experiences, including graduating top of my class at Cal after navigating homelessness, gave me a unique perspective and a drive to build useful, ethical tools. I specialize in bridging the gap between 'notebook data science' and scalable infrastructure. My core expertise lies in architecting end-to-end data systems—from raw ingestion (Dagster/dbt) and warehousing (BigQuery/Athena) to deploying predictive models (Propensity Scoring, NLP, Churn) that drive revenue. I bring a software engineering mindset to data teams, championing CI/CD, unit testing, and modular design. Whether acting as a solo data lead or a core contributor in high-growth startups, I focus on untangling complexity to build tools that are explainable, maintainable, and directly impactful.

SKILLS

Programming & Core Data Skills	Machine Learning - Predictive & Classical	Deep Learning & Advanced AI
Python (Pandas, NumPy, SciPy)	Propensity Scoring	Deep Learning (PyTorch, TF/Keras)
SQL (PostgreSQL, BigQuery) R	Churn Modeling	Neural Nets (CNN, RNN, Transformers)
Bash/Shell Scripting	Predictive Modeling	Reinforcement Learning Concepts
Excel for Analysis	Predictive Risk Modeling	
Statistical Analysis	Classification & Regression	
Feature Engineering	Clustering Algorithms	
Unstructured Data Analysis	Recommender Systems	
	Bayesian Methods (PyMC3)	
	ML Frameworks (Scikit-learn)	
	Model Eval & Selection	
	Time Series Analysis/Forecasting	
Natural Language Processing (NLP)	Generative AI & LLMs	Data Engineering & Cloud Platforms
NLP (Sentiment, Topic Models)	Prompt Engineering	AWS Cloud (EC2, S3, Lambda, Athena)
Text Vectorization (TF-IDF, W2V)	LLM Fine-tuning	GCP Cloud (BigQuery, GCS)
NLP Libs (spaCy, NLTK, HuggingFace)	Retrieval Augmented Generation	Data Pipelines (Dagster, dbt, dlt)
Info Extraction & Entity Recog.	LangChain LangGraph	Data Warehousing (BigQuery, Athena)
	Agentic AI	Database Modeling (Star, Snowflake)
	LLM Evaluation & Benchmarking	ETL/ELT Design & Implementation
	Open-Weights Models (Llama)	API Design & Integration
		Docker & Containerization

Software Engineering & DevOps Practices

- API Utilization
- Production Code Quality/Practices
- Object-Oriented Programming (OOP)
- Software Design Patterns
- Testing (Unit, Intg, E2E, Pytest)
- CI/CD Pipelines (GitHub Actions)
- Version Control (Git, GitHub)
- Agile Methods (Scrum, Kanban)
- Code Review & Collaboration

Collaboration & Professional Skills

- Team Leadership & Mentoring
- Cross-functional Collaboration
- Stakeholder Comms & Management
- Requirements Elicitation/Define
- Project Management (Agile)
- Technical Documentation/Reporting
- Presenting (Tech & Non-Tech)
- Problem Solving & Critical Think

Familiar Technologies & Other Tools

- Java
- JavaScript
- HTML/CSS
- C++ (basics)
- PHP (basics)
- Databases (PostgreSQL, MySQL)
- Web Scraping (Beautiful Soup, Selenium)
- Apache Spark (PySpark basics)
- Jupyter Notebooks & Lab

Data Visualization & BI Tools

- Plotly (Dash for web apps)
- Tableau
- Looker Studio
- Matplotlib & Seaborn Graphics
- Geospatial Viz (ArcGIS, GeoPandas)
- Interactive Dashboard Design
- Business Intelligence Reporting

MLOps Principles & Tools

- Explainable ML
- Experimental Design (A/B)
- Causal Inference Methods
- Quantitative & Qualitative Data
- Survey Design & Analysis
- Statistical Significance Tests
- Ethical AI & Data Bias Analysis
- Data Privacy & Security

Research, Experimentation & Ethics

WORK EXPERIENCE (4)

Sep 2024 - Oct 2025

Data Scientist (Lead: ML, Data Engineering, MLOps) - Ed Pioneers Fellow at Caliber Public Schools

Richmond, California

- Strategic Leadership (Solo Data Lead):** Owned the full data lifecycle (DS, DE, ML) as the sole data scientist. Partnered directly with C-suite and department heads to navigate a 'zero-to-one' environment, moving the org from manual spreadsheets to automated warehousing.
- Predictive ML & Risk Modeling:** Developed and deployed explainable ML models (Logistic Regression, Deep Learning) to predict staff turnover. Engineered a 'Risk Tolerance' configuration allowing non-technical leadership to adjust precision/recall thresholds based on quarterly hiring capacity.
- Modern Data Stack Architecture:** Architected and built a scalable platform on Google Cloud Platform (GCP). Orchestrated ELT pipelines using Dagster, dbt, and dlt to ingest data from disparate SIS (School Information Systems) and HR platforms.
- Engineering Maturity (ROI):** Engineered a comprehensive People Team data pipeline, reducing manual consistency checks from months of collective annual work to seconds. Built a custom automated data quality framework to catch inconsistencies between systems.

- **API Integration & Interoperability:** Solved complex data interoperability challenges by designing integrations between Schoolmint and PowerSchool. Reverse-engineered undocumented APIs to create a unified data model for cross-functional analysis.
- **Data Democratization:** Updated the organization's data security policy and designed data literacy training modules, empowering school leaders to access real-time attendance and academic metrics without technical bottlenecks.
- **Survey Design & Enrichment:** Leveraged data to enrich staff satisfaction surveys with demographic, work location, role, grade level, and tenure data, enabling highly segmented and actionable insights.
- **Stakeholder Management:** Partnered with leadership on high-impact, data-driven solutions; presented findings to stakeholders including the Board of Directors.

Aug 2021 - Feb 2023

Data Scientist (ML, Data Engineering, MLOps) at SetSail

📍 San Mateo, California

- **Business Impact:** Contributed to product enhancements that achieved 33% faster ramp times, 16% higher revenue, and 15x ROI for customers.
- **Production ML (Revenue):** Developed and deployed production ML models for Propensity Scoring (deal closure probability) and Churn Modeling. Leveraged NLP on unstructured email metadata to identify sales signals.
- **Pipeline Architecture (AWS):** Led a critical overhaul of the AWS data infrastructure (S3, Athena, EMR). Implemented 'SQL Push-down' strategies and asynchronous DAGs, reducing data processing latency by 75% and scaling to multi-terabyte datasets.
- **Scalable Data Modeling:** Architected scalable Star Schema data models and optimized ETL/ELT processes, ensuring data readiness for LLM integration.
- **Engineering Best Practices:** Championed the adoption of CI/CD pipelines (GitHub Actions), unit testing (pytest), and Agile methodologies within the data science team.
- **Causal Analysis:** Performed deep causal inference studies to isolate specific sales behaviors that drive outcomes, influencing the product roadmap to focus on 'high-leverage' user actions.
- **Technical Consulting:** Acted as a technical consultant for enterprise customers, diagnosing complex data discrepancies and proposing architectural solutions for data integration.
- **Cross-Functional Leadership:** Collaborated seamlessly with Engineering, Product, and Support teams to spec out large-scale infrastructure restructuring.

Sep 2020 - May 2021

Data Science Research Team Lead at UC Berkeley School of Public Health

📍 Berkeley, California

- **Leadership:** Led data science components for mixed-methods studies on equity and public health (specifically violence against homeless youth). Managed a team of undergraduates.
- **Unstructured Data:** Analyzed diverse unstructured and non-traditional datasets (qualitative interviews, geospatial data, text corpora, hand-drawn maps) requiring the development of novel data processing approaches.
- **Geospatial Analysis:** Performed geospatial analysis (ArcGIS) to identify and visualize spatial violence patterns for non-technical stakeholders.
- **Visualization:** Created interactive dashboards (Tableau, Plotly) to communicate findings to stakeholders.
- **Interdisciplinary Collaboration:** Collaborated across disciplines (public health, psychology, sociology) ensuring ethical, robust research.

May 2019 - Aug 2019

Full Stack Engineer at Los Medanos College

📍 Pittsburg, California

- **Full Stack Development:** Independently designed and developed a web application (Python, JavaScript, PHP, PostgreSQL) to guide student program selection.
- **Data Acquisition:** Built a database of 500+ university programs via web scraping (Beautiful Soup) and rigorous data cleaning.
- **Stakeholder Consulting:** Consulted with college stakeholders (district, counselors) to define user needs.
- **UX/UI:** Created an intuitive management dashboard for non-technical staff.

PROJECTS (5)

ResumeLLM (AI Agent for Resume Tailoring)

↗ <https://github.com/jstehn/resume-llm/tree/develop/>

Python Generative AI LLMs LangChain LangGraph Agentic AI NLP Data Optimization

- **Agentic AI System:** Developed a robust, agent-based system for precise and context-aware resume customization based on specific job descriptions and user preferences.
- **Complex Orchestration:** Leveraged LangChain and LangGraph to design and implement sophisticated AI agent orchestration, enabling complex multi-step reasoning and dynamic workflow management.
- **NLP & Parsing:** Applied advanced Natural Language Processing (NLP) and Large Language Model (LLM) techniques to parse job requirements, extract key skills, and generate tailored resume content.
- **Iterative Optimization:** Designed the system to iteratively refine resume drafts, demonstrating capabilities in iterative AI feedback loops and intelligent document optimization.

Dagster People Team Pipeline

↗ <https://github.com/jstehn/dagster-people-team-pipeline>

Python Dagster Data Pipeline ETL GCP BigQuery Data Engineering

- **Automated ETL:** Architected and implemented a scalable ETL pipeline using Dagster for orchestrating data workflows from various HR sources (BambooHR, Paycom, internal docs) into Google BigQuery.
- **Automated QA:** Automated data consistency checks, reducing manual validation efforts from months annually to seconds, significantly improving data accuracy and reducing human error across HR systems.
- **Predictive Analytics:** Enabled rich longitudinal and cross-referential data analysis, enriching staff satisfaction surveys with detailed segmentation (location, role, demographics, tenure).
- **Data Interoperability:** Designed the pipeline with explicit consideration for integration with diverse data systems and platforms, ensuring broad interoperability.

Firewatch Chatbot (CalSAFE DataGood)

↗ <https://github.com/jstehn/firewatch-chatbot>

Python Chatbot NLP Intent Classification spaCy scikit-learn Flask Heroku Facebook API DataForGood Jupyter

- **Intent Classification:** Built NLP intent classification model using spaCy (vectorization) & scikit-learn (Logistic Regression).
- **Dataset Creation:** Manually collected & classified ~450 question dataset with team.
- **Bot Logic:** Designed bot logic providing responses based on classified intent.
- **Deployment:** Integrated with Facebook API & planned deployment via Flask/Heroku.

Warren Court Language Model

 <https://github.com/jstehn/warren-court-language-model>

Python NLP Language Modeling Jupyter Data Analysis History

- **Language Modeling:** Applied language modeling to analyze patterns in historical legal text.
- **NLP Exploration:** Focused on personal learning and exploration of NLP techniques.

Neural Network By Hand (Berkeley Exercise)

 <https://github.com/jstehn/nn-by-hand-188>

Python Neural Networks Machine Learning NumPy Education

- **Core Implementation:** Implemented core NN components (layers, activation, backpropagation) based on class skeleton.
- **Mathematical Foundations:** Focused on understanding mathematical foundations over code flexibility.
- **Deep Learning Concepts:** Solidified grasp of fundamental deep learning concepts.

VOLUNTEER

Sep 2024 - Sep 2025

Impact Fellow (Placement @ Caliber Public Schools) at Education Pioneers

Selected for national fellowship applying leadership/management skills to advance educational equity via capacity-building projects and leadership development.

- **Leadership Development:** Applying data science & leadership skills to advance educational equity.
- **Capacity Building:** Building organizational capacity through strategic data projects at placement site.
- **Cohort Engagement:** Engaging in rigorous leadership development programming with diverse cohort.

Dec 2023 - Current

Data Team Lead at San Francisco Gay Men's Chorus

Provide data-driven insights for policy-making and organizational growth through survey creation and analysis (qualitative & quantitative).

- **Team Leadership:** Led volunteer team providing data analysis for organizational strategy.
- **Survey Analysis:** Designed & analyzed surveys (qual/quant) informing policy & growth.
- **Executive Presentation:** Presented data-driven insights to chorus leadership.

Event Producer at Bearrison Street Fair

- **Event Production:** Co-produced large-scale (~10k attendees) LGBTQ+ community street fair, overseeing all aspects from planning through execution.
- **Logistics Management:** Managed complex logistics, ~100 vendor relations, and multi-stage entertainment programming.
- **Fundraising:** Led fundraising efforts, securing over \$90k in sponsorships & donations, contributing to event profitability.
- **Stakeholder Coordination:** Coordinated hundreds of diverse stakeholders including volunteers, performers, operations teams, city agencies, and non-profits.

Jan 2018 - Current

Mentor and Trans Support Leader at San Francisco Gay Men's Chorus

Support members through mentorship and leadership within trans support initiatives, coordinating meetings and events.

- **Peer Mentorship:** Provided mentorship and peer support to chorus members.

- **Community Leadership:** Led coordination for trans member support group meetings and events.
- **Inclusive Culture:** Contributed to fostering an inclusive environment within the organization.

Transfer Mentor at UC Berkeley Division of Computing, Data Science, and Society

- **Student Mentorship:** Mentored incoming transfer students transitioning into UC Berkeley Data Science.
- **Academic Guidance:** Assisted students in developing data science skills & navigating coursework.
- **Community Building:** Fostered community and peer networking during remote learning (pandemic).

Jul 2017 - May 2019

Student Ambassador (Transfer & Career Services) at Los Medanos College

Supported transfer/career programs through data analysis, marketing, peer training, and event coordination.

- **Data Analysis:** Analyzed student transfer data (SQL, R, Excel) to inform program development.
- **Marketing Leadership:** Led marketing committee managing social media, web content, and outreach.
- **Public Speaking:** Presented transfer/career information via public speaking & workshops.
- **Peer Training:** Trained new student employees on department policies & procedures.
- **Event Coordination:** Organized large campus events coordinating multiple stakeholders.

EDUCATION (1)

Aug 2019 - May 2021

Bachelor of Arts Data Science (Domain Emphasis: Quantitative Social Science) at University of California, Berkeley

COMPSCI C8: Foundation of Data Science COMPSCI 61A: Structure & Interpretation of Computer Programs

COMPSCI 61B: Data Structures COMPSCI 188: Artificial Intelligence DATA C100: Principles & Techniques of Data Science

DATA C102: Data, Inference, and Decisions STAT 134: Concepts of Probability STAT 153: Time Series Analysis

STAT 89A: Linear Algebra for Data Science DEMOG 110: Population Analysis HISTORY C184D: Ethics of Data

Grade: 4.00/4.00

Highest Distinction (Summa cum laude). Outstanding Data Science Undergraduate Award (Top of Class).

AWARDS (1)

May 2021

2020-2021 Outstanding Data Science Undergraduate Award at UC Berkeley

Recognized for excellence in Data Science undergraduate studies, research, and community contributions at UC Berkeley.

INTERESTS (2)

Technology & Tinkering Creative & Community

AI Research & Trends

Singing

Playing Instruments

Generative AI (Local Models)

Dancing

Language Learning

Large Language Models (LLMs)

Community Engagement

Raspberry Pi Projects

REFERENCES

"Chosen from over 50 applicants and 5 finalists, Jack joined our organization at a pivotal moment and has been an invaluable team member ever since. As soon as they joined, they immediately took initiative on a complex survey design and analysis project that was critical to our success, bringing both expertise and ownership from day one. Jack's approach is highly collaborative and mission-driven. They actively engage with departments across the organization, listen closely to

their needs, and build thoughtful, scalable solutions—including dashboards, data quality reports, and automated systems that allow staff to focus on their core work with students. Jack is systems-oriented and consistently plans for long-term, sustainable outcomes. One standout example: Jack streamlined a survey and analysis process that previously took our team a month, developing a replicable system that now delivers actionable insights in just a few days. This perfectly captures their ability to problem-solve proactively and significantly boost our efficiency and decision-making. Beyond their technical and strategic skills, Jack is reliable, resourceful, and generous with their knowledge. They've led internal trainings to empower colleagues, handle ambiguity with ease, and bring a positive, solution-focused mindset to every challenge. Jack is an easy choice for any team seeking a results-driven, collaborative data scientist who elevates both projects and people. I recommend them without hesitation. ”

Brian Jimenez (Managed Jack directly at Caliber Public Schools) - Managing Director of People

“ I had the pleasure of interviewing Jack before they joined the SetSail team-- I gave them a 4 out of 4. It's important to note that on our hiring scale, a 4 meant "I will flip the table if you don't hire this person." One thing that stuck with me after the interview, and which was reaffirmed while we worked together at SetSail, is Jack's enthusiasm for data science and their love of learning (and sharing what is learned). Not only is Jack an extremely capable engineer and data scientist, they are also a collaborative team player who elevates everyone around them. Their contributions at SetSail were always valuable to the company-- whether it was their huge role in our data pipeline migration, or countless bug fixes and feature implementations that directly improved our user experience, you could always count on Jack to get the job done on time, with clean code, and great documentation. I wholeheartedly recommend Jack for any data science position—they would be an invaluable addition to any team. ”

Darrin Gilkerson (Worked with Jack on different teams at SetSail) - Software Engineer at QVT Financial

“ Jack is a sharp, human-first data person. They possess incredible passion for doing what is right and making good science happen. I highly recommend their work and their presence. ”

Ollie Downs (Studied with Jack at UC Berkeley) - Senior Data and Research Analyst, County of San Diego

“ Jack was an integral part of the planning and designing of data pipeline overhaul at SetSail. Even with a moving target and many dependencies, Jack was able to adjust the design of our new pipeline, maintaining conversations across the product and engineering teams as the project progressed. They are also a fast learner and willing to dig into new technologies, which I really admired as their coworker. They would be a great addition to any team looking for a fast-learning and flexible data scientist. ”

Sarah Nam (Worked with Jack on the same team at SetSail) - Senior Associate at Cancer Navigator

“ Jack is a hard-working Data Scientist with a keen eye for details. Their passion for data analytics and software development really stands out when tasked with complex problems. At SetSail, Jack worked on a variety of projects that involved teasing out actionable insights from complex data sets, enhancing modeling capabilities through feature development and algorithm development, and building out a data ETL process that transformed the data infrastructure to help SetSail scale for enterprise customer needs. In addition to these technical skills, Jack's collaborative work with the engineering and product team continually earned praises from fellow coworkers. They were never shy and was always proactive to jump in and help solve a problem. I highly recommend Jack as a Data Scientist and Data Engineer for any organization. Their technical skill and work ethic will be immediately apparent upon joining any team. Feel free to reach out to me as I am happy to provide additional reference or information as desired. ”

Danny Pan (Managed Jack directly at SetSail) - Data Science

“ I had the privilege of working alongside Jack at SetSail, and I can confidently say that they are a top-notch data scientist. Jack's expertise in data science, combined with their passion for software engineering, make them a valuable asset to any team. They have a keen ability to plan and lead complex cross functional projects and their software engineering skills are second to none. Jack's enthusiasm for learning is contagious, and they are always eager to dive into new projects and technologies. They are a great communicator and are able to explain technical concepts in a way that is easy for both technical and non-technical colleagues to understand. On top of all that, Jack is one of the kindest and most genuine people I've had the pleasure of working with. They truly care about their team and go above and beyond to support them. I highly recommend Jack for any data science or software developer role, and I have no doubt that they will excel in their next endeavor. ”

Josh Mantovani, M.A. (Senior to Jack, worked together at SetSail) - Data Scientist / Engineer

“ I am happy to recommend Jack for a variety of roles and positions. Jack is a motivated self-starter who loves to accomplish project tasks while developing and implementing smooth processes in their work environments. Jack is an accomplished leader, utilizing problem-solving skills to support their own work and the work of their colleagues and peers, taking time to ensure that their team has the skills, knowledge, and resources they need to finish their tasks and projects effectively. Jack has a wide array of skills that they readily apply to their work, and they are ready to search for answers and learn new skills to address problems that arise in their projects. Then they are ready and willing to teach peers and colleagues how to utilize those new skills, supporting team-based processes and accomplishing team projects and goals in addition to their own individual work. Jack is a leader who uses imagination, experience, and empathy to create sustainable processes and consistently complete their goals. I am happy to recommend Jack and I am confident that Jack will be a positive asset to any work that they set out to complete. ”

G. Allen Ratliff (Managed Jack directly at UC Berkeley SPH) - Assistant Professor of Social Work

“ I have had the great fortune of having Jack as project lead on the SFYEAH Research Project. One word to describe Jack, I would say "Integrity", Jack holds themselves to the highest standard. It shows in the work they produce, Jack is meticulous. Jack is skilled coder and data scientist, with a wealth of geospatial analysis knowledge. They are are first rate leader, and an exceptional communicator. Jack keeps everyone on the same page, and is incredibly thorough. It is an absolute pleasure to work with Jack. ”

Conan Minihan (Jack was Project Lead for SFYEAH Research Project) - Data Scientist, PhD Student

“ Jack and I worked on the same research team and they effortlessly evolved into a pillar of leadership and direction. It's been an absolute pleasure and relief to be able to work alongside them. Jack learned quickly and worked beyond the expected and required amount to ensure deadlines and quality were kept. It truly astounded me how enthusiastic and exceptionally intelligent Jack was as I watched them surpass most of the team in their domain of expertise and knowledge in a matter of weeks. Jack's passion for details, design, and accuracy has made them one of the strongest assets on our team. Their work ethic and energy have and are contagiously inspiring and addicting to be around. Jack is just one of those people that you want on your team in every scenario because they really own the title "jack of all trades" ”

Eva Smolentseva (Worked with Jack on same research team at UC Berkeley) - Analyzing Natural Language Models @USAA