Email | jeremy@thaller.dev
GitHub | github.com/jthaller
Portfolio | jeremythaller.com

JEREMY THALLER

Brooklyn, NY

Senior full-stack data scientist with two MS degrees and experience productionalizing ML models in fintech and logistics

EXPERIENCE

VEHO | Last-mile delivery company powered by a dynamic driver marketplace
Senior Data Scientist, Marketplace | *New York, NY [May. 2022 – Present]*

- Revamped route generation algorithm, leveraging simulations to optimize variables and implemented changes that yielded a substantial 17 cents per package cost reduction, translating to over $2M in annual savings.
- Implemented stateless supervised and unsupervised ML algorithms within the route-generating microservice
- Refactored external API calls to spin through a rate-limiting backoff loop and log success metrics in new relic

CURRENT | Financial technology company including no-fee debit banking, debit cards, early payroll deposits, and crypto trading
Data Scientist, Core-Banking / Product | *New York, NY [Aug. 2021 – May 2022]*

- Achieved 95% accuracy in predicting long-term, user-level segmentation from 30d of user behavior; segments were assigned via k-means clustering and PCA, and the early indicators model was trained with Vertex AI and BigQuery ML; resulting model provided feedback for marketing spend optimization in 1/6 the time as previously required.
- Modeled different product gating criteria to forecast potential churn and revenue scenarios; implementation resulted in $200k/month in cost savings, as estimated through causal impact analysis determined with SARIMA forecasting.
- Wrote DAGs and SQL queries to populate daily updating BQ tables powering data studio dashboards; later contributed to the DBT migration from Airflow

CELSIUS NETWORK | Centralized finance crypto company providing comprehensive financial services, formerly with $30B AUM
Data Analyst, Growth and Marketing | *New York, NY [Sept. 2021 – Aug. 2022]*

- Utilized supervised machine learning to identify potential accredited investors with 98.4% accuracy; expanded the number of high net worth prospects by 10x, delivering $50B of potential platform growth
- Determined feature importance for third party prospecting via unsupervised machine learning (clustering)
- Designed, developed, and deployed an ML anomaly detection algorithm to alert fraud and platform violations
- Developed internal Python packages to standardize and expedite repeated SQL queries and data transformations
- Initiated data augmentation through user surveys; utilized qualitative and quantitative results—often with Bayesian statistics—to predict vectors of user and asset growth, as well as steer brand image and marketing strategies
- Developed and presented weekly research to executive stakeholders, steering strategic GTM initiatives

BROOKHAVEN NATIONAL LABORATORY | Structure and Dynamics of Applied Nanomaterials
MS Thesis Researcher in Deep Learning | *Upton, NY [Feb. – Sept. 2021]*

- Reduced simulation compute time by 50x by developing a new statistical-based methodology
- Utilized TensorFlow to predict absorption spectra disorder of Au nanoparticles, reducing data required by 90%
- Created and managed lab's GitHub organization; constructed example projects to demonstrate best dev practices

EDUCATION

LUDWIG MAXIMILIANS UNIVERSITÄT, Munich, Germany | 2019 – 2021
M.S. Materials Science and Engineering
ADAM MICKIEWICZ UNIVERSITY, Poznań, Poland | 2019 – 2021
M.S. Computational and Applied Physics
WILLIAMS COLLEGE, Williamstown, MA | 2015 – 2019
B.A. Physics, Honors | Sigma Xi Honor Society | Varsity Track & Field Captain

SKILLS AND TOOLS

Programming Languages (Years of Experience) | Python (5), SQL (3), Java (7), MATLAB (4), R (1), Julia (1)
Python Packages | Pandas, NumPy, Scikit-Learn, Numba, PyTorch, TensorFlow, Keras, PySpark, Regex, WandB, Dask
Data Visualization Software | Looker, Data Studio, Excel/GSheets, Mathematica, Jupyter Notebooks, WandB, Plotly
Data Engineering Tools | Snowflake, Apache Airflow, Docker, PySpark, DBT, AWS

Email | jeremy@thaller.dev
GitHub | github.com/jthaller
Portfolio | jeremythaller.com

INDEPENDENT PROJECTS

SPOTIFY ETL AND RECOMMENDATION ALGORITHM

- Leveraged PySpark, scalar-aggregate-reduction optimized SQL queries, & the Spotify Web API to investigate song trends as well as song/genre characteristics via dimensionality-reduction and cosine distance
- Trained a recommendation algorithm using song embeddings trained via Gensim's Word2Vec on 1M Spotify playlists
- Setup an airflow DAG to extract my daily listening history the via Spotify Web API and write it into a personal Postgres Database in a docker container. Then, leveraged the recommendation model to push an AI-recommended playlist to my personal Spotify account each week

CRYPTOCURRENCY EDA AND PREDICTIVE MODELING

- Analyzed DOGE Coin value and key financial indicators over time with interactive plots
- Forecast DOGE Coin values in Python using Keras and an LSTM architecture

Fast-paced environment
Deadline-driven
Multi-faceted
Scraping→Producing→Presenting→Interpreting→Guiding business decisions
Liaising

Fix spacing between bullets
Find replacements for utilization

- Fulfilled high-stake data request under time pressure

CRYPTOCURRENCY EDA AND PREDICTIVE MODELING

- Analyzed DOGE Coin value and key financial indicators over time with interactive plots
- Forecast DOGE Coin values in Python using Keras and an LSTM architecture

***Tell us about an analysis, data presentation, or dashboard you put together, and how this analysis contributed to the success of your business.**

Brief context:
Celsius (my employer), a crypto company recently issued regulations, limiting their major product to only be eligible to "accredited investors" (AI). Accordingly, it became a top strategic priority to convert as many users to AI as possible and track this progress.

The work:
In short, I took ownership of all AI analytics, from data engineering to executive presentations. To track the progress for AI conversion, I defined a series of KPIs and created a dashboard in Google Sheets, the highlight of which was a tree diagram to easily visualize the progress of different segments across the business. To update the the sheet, I wrote a python script with Postgres SQL queries to pull the relevant data. I then added a CRON job to run the script daily, and utilized the GSheets API to populate the dashboard with the updated data.

The AI dashboard became the de-facto truth source for the initiative, and I used it to walk through the progress and current state of the business with multiple executives. After providing enablement to product managers, they were able to effectively maintain and lead their own business analytics, thus increasing overall data literacy. I then shifted my focus on providing a list of highly likely candidates for accreditation, a process requiring supervised machine learning.

Dear Psychology Today,

I am currently a data analyst in Growth and Marketing at Celsius Network – a major CeFi (centralized finance) crypto lending platform. My passion is focused on utilizing data and statistical methods to uncover

Email | jeremy@thaller.dev
GitHub | github.com/jthaller
Portfolio | jeremythaller.com

insights, build sophisticated predictive models, and help inform strategic decision making. I am writing to introduce myself and express my interest for the Data and Business Intelligence Analyst role at Psychology Today.

Psychology Today's position as the most visited psychology website is exciting, but not an excuse for complacency. A broad understanding of the competitive landscape, as well as user sentiment through social listening and in-house web scraping and NLP are two aspects which I see as critical in maintaining market dominance and improving the website and magazine.

While being a member of the data team at Celsius, I also fully embedded myself in the client-facing side of the business. Far too often, data analysts and scientists limit themselves to only dealing with the data, and stopping short of seeing the full impact of their research. By presenting my insights and working closely with the go-to-market teams, I was able to gain a unique insight among the team; combining the unique visibility I have in data alongside my direct business insights has allowed me to provide more holistic program recommendations.

Conversely, my embeddedness has allowed me to
combine the unique visibility I have in data alongside business insights to give often more holistic program recommendation

I gained a full-stack understanding of the imp and implementations of my data insights. While at Celsius, I used a third party vendor (MeltWater) to track social listening and user sentiment over time, especially in relation to company scandals and industry-wide "FUD." To understand the impact of our sponsored influencers, I scraped audio from their youtube videos, translated it to text, and analyzed the text with various NLP techniques (LDA, sentiment distribution analysis, term-frequency and TF-IDF analysis). To understand the user lifecycle journey, specific sub-product behaviors, and our changing demographic of registrants, I led surveys and presented my findings, which altered our marketing strategies and messaging.

I led multiple user surveys to determine potential areas of asset growth, specific product use, and better understand user lifetime journey.

I am a dual master's thesis candidate concentrating in machine learning driven materials science and engineering

Although my background is in academic research pertaining to physics, I am seeking to apply my strong statistics and programming experience to the business setting. During my first semester of graduate studies in Munich, I took a course on computational materials design where I had my first exposure to machine learning. During the course, I developed my skills in data science and built a foundation of predictive modeling. The course culminated in a final project where I utilized principal component analysis and distributed processing to deal with high-dimensionality and big data, applying machine learning to accelerate potential semi-conductor screening. Throughout the process, I saw the implications data science techniques have for commercial problems that I wanted to solve. Since then, I refocused my remaining studies on deep learning and Bayesian statistics for the purposes of helping data-centric businesses develop increasingly sophisticated means by which to improve holistic customer experiences and meet KPI's.

In my free time, I tackle projects that involve the end-to-end stages of the data science process. For example, after scraping my Facebook Messenger data via selenium and beautiful soup, I've been building increasingly complex natural language processing (NLP) models and analyzing my messaging trends. A recent project of mine trains a model by ingesting organic Facebook Messenger data, creating a naïve Bayes classifier to model unlabeled messages, continuously improving its predictive power as more data is added. I'm now utilizing transfer learning to build a robust chatbot, capable of conversing with my voice. Currently, I've been exploring a massive Spotify dataset on Kaggle, using optimized PySpark SQL queries to explore how user listening patterns have changed over time. In my current thesis work at Brookhaven National Lab, I ran complex simulations on the BNL distributed computing cluster and analyzed the resulting data in Python using pandas dataframes, PySpark, matplotlib/seaborn, and Plotly. Using these simulations as training data, I built neural network with TensorFlow and applied transfer learning to decode the complex structural information of nanoparticles from particle accelerator data. As such, I believe that the complexity and range of my project experience prepares me well to uncover the novel insights and

Email | jeremy@thaller.dev
GitHub | github.com/jthaller
Portfolio | jeremythaller.com

modeling needs McKinsey's partners will require. Core problems requiring data/text mining, NLP, user-profiling, and machine learning are exactly the types of problems I can work together with the McKinsey's partners to solve.

To close, I believe my strong analytical background and experience in managing complex data science problems involving NLP and big data will make me a valuable addition to McKinsey. Each Friday for the last six-months, I've presented the progress I've made on my deep learning project. Over this time, I've built a strong foundation of deconstructing highly technical concepts for my boss and coworkers, many of whom have little exposure to ML. If selected, I would make it my goal to help foster a strong team culture, build cutting edge models, and clearly explain our key insights and methodologies to organizational stakeholders. Thank you for your time and consideration. I hope to hear from you soon.

Sincerely,

I have two master's degrees in materials science and engineering. During my university studies, I focused on computational materials design via deep learning, culminating in my Msci thesis where I utilized simulations, keras, and transfer learning.

In my current FinTech role, I've implemented machine learning to detect fraud, understand our customers with unsupervised ML, and predict the probability for any user to be eligible for accreditation. This last project, which utilized supervised ML, involved purchasing data from a third-party vendor, a lot of EDA, data engineering, pipelines, and hyperparameter tuning), as well as collaborating with DevOps to create a database schema that would work for the ML-Ops process of continually improving the model with updated information.

I've been coding in python nearly every day for the last 3 years, and I had learned the language two years prior to that. Python is my preferred language, and when I think of data structures and algorithms, I think of their python version first. I've implemented some highly optimized programs using numba's just-in-time complication, multi-threading, and Cython (c python) implementations of pandas dataframe transformations.

As for R, I am familiar and have used it to calculate some basic statistics in cases when it's easier to do in R than Python, but in general, I avoid R in favor of Python.

Old Celsius full bullet points

CELSIUS NETWORK | Centralized finance crypto company providing comprehensive financial services with $20B+ AUM
Data Analyst, Growth and Marketing | *New York, NY [Sept. 2021 – Aug. 2022]*

- Utilized supervised machine learning to identify potential accredited investors with 98.4% accuracy; expanded the number of high net worth prospects by 10x, delivering $50B of potential platform growth
- Determined feature importance for third party prospecting via unsupervised machine learning (clustering)
- Designed, developed, and deployed an ML anomaly detection algorithm to alert fraud and platform violations
- Developed internal Python packages to standardize and expedite repeated SQL queries and data transformations
- Initiated data augmentation through user surveys; utilized qualitative and quantitative results—often with Bayesian statistics—to predict vectors of user and asset growth, as well as steer brand image and marketing strategies
- Developed and presented weekly research to executive stakeholders, steering strategic GTM initiatives
- Analyzed sponsored influencer effectiveness and changed behavior using speech to text APIs and NLP techniques
- Oversaw the secure data exchange with third party vendors, using an AWS S3 bucket & best encryption practices
- Automated and built dashboards via python, GSheets API, and BASH scripts; later migrated these dashboards to Looker
- Contributed to database snowflake migration with DBT, new materialized views, and Apache Airflow DAGs
- Defined new KPI metrics to better track user and platform growth via marketing efforts, irrespective of externalities
-

YALE UNIVERSITY | Mechanical Engineering / Materials Science
Researcher – Solid State Physics and Metallurgy | New Haven, CT *[Summer 2019]*

- Wrote and deployed a GUI Python program to automate and expedite material candidate screening
- Formulated an experiment to isolate the causal variable behind thermo-mechanically molded nanowire orientations

FACEBOOK MESSENGER ANALYSIS

- Scraped 10+ years of messaging data via selenium and BS4; analyzed messaging trends with Pandas, NLTK, SpaCy, and Gensim, showcasing the results with charts and word clouds
- Created a 'friend' classifier through Bayesian statistics, capable of predicting which friend sent an unseen message
- Built a from-scratch generative chatbot trained on personal messaging data using Keras and GloVe embeddings