# Alexandre Matton

*Machine Learning Research Engineer in New York City*

USA: +1 650 334 9206
matton.alex@gmail.com
https://github.com/alex-matton

## Education

2018–2020 **Stanford University, California, USA**.
- MSc in Computational and Mathematical Engineering (ICME), GPA: 4.16 (Top 5%)
- Relevant coursework: deep learning in NLP, statistics/probability/classical ML, RL, advanced linear algebra, recommendation systems, mining of massive datasets

2015–2018 **École Polytechnique, Paris, France**, *France's best engineering school according to the Times Higher Education rankings*.
- MSc in computer science (ML and theoretical) and applied mathematics, GPA: 3.9/4.0 (Top 10%)
- Ranked 5th at the entrance examination, among more than 7,000 students

2013–2015 **Collège Stanislas, Paris, France**, *Post-secondary course leading to nationwide competitive entrance exams to the French Grandes Écoles for scientific studies*.
- Mathematics, computer science and physics, GPA: 3.98/4.00 (Top 5%)

## Experience

10/2022 - current **Member of Technical Staff, Cohere AI**.
- Developed an automated way to train text classification models on very low amount of data with performance on par with state-of-the-art. The new training pipeline is an order of magnitude faster than open-source alternatives
- Built a pipeline to finetune semantic search models providing superior performance than the current state-of-the-art on popular adademics benchmarks (BEIR)
- Built a highly scalable deployment infrastructure with very high throughput and extremely low cold boot time

08/2020 - 10/2022 **Senior Machine Learning Research Engineer, Scale AI**.
- Developed an end-to-end ML platform that reads and extracts information from documents such as receipts and invoices. Worked on all aspects of the pipeline (labelling optimization, modeling, deployment, monitoring). Headed R&D NLP efforts. The platform can process 1M+ pages per day and returns results in $\leq$ 3s/page.
- ML lead for an RFP/demo that we won against 10+ competitors which led to a multi-million dollar deal
- Acted as temporary tech lead and ran a team of 5 engineers
- Conducted 150+ interviews and participated in the design of contracts with new customers
- Fastest ML engineer to get promoted from junior to senior

03/2019 - 03/2020 **Teaching Assistant, Stanford University**.
- Advanced graduate-level course in Numerical Linear Algebra taught by Prof. Darve (CME302), and Deep Learning for NLP taught by Prof. Manning (CS224N)
- Undergraduate-level course in Linear Algebra and PDE's taught by Prof. Khayms (CME104)

06/2019 - 09/2019 **Machine Learning Engineer Intern, Twilio**.
- Implemented BERT-like models as services in the ML platform
- Analyzed the redundancy and location of information contained in output vectors of such models
- Designed unsupervised algorithms to extract quality phrases from a corpus of text using statistical methods

04/2018 - 08/2018 **Research Intern in crypto-currencies, S.R.I. International**.
- Studied general behaviors of Bitcoin owners, with a stress on the ones undertaking illegal activities
- Designed heuristics for NP-complete problems to track specific types of trades in blockchains

## Papers and Blog Posts

2021 **DEBAGREEMENT: A comment-reply dataset for (dis)agreement detection in online debates**, *NeurIPS 2021 Datasets and Benchmarks Track*.

2020 **A Survey of Deep Learning Approaches for OCR and Document Understanding**, *34th NeurIPS 2020 Workshop: ML Retrospectives, Surveys & Meta-analyses (ML-RSA)*.

2019 **Emergent Properties of Finetuned Language Representation Models**, *On arXiv, written while at Twilio.*

2019 **Longitudinal Analysis of Misuse of Bitcoin**, *ACNS2019, written while visiting the S.R.I.*

2020 **Fast transformer decoding via caching**, *https://scale.com/blog/pytorch-improvements*.
Blog post on transformers' decoding speed optimization (between 40% and 250% faster than the PyTorch implementation).

## Skills

Programming Python (Pytorch, PySpark, transformers, sklearn, FastAPI), Go, Java, C++, Julia, OCaml

Tools Kubernetes, Docker, Triton, WandB & MLFlow, SageMaker, AWS & Gcloud, Git

Languages French: Native, English: Fluent, Spanish: Intermediate