Emma Scharfmann

https://github.com/EmmaScharfmann
https://emmascharfmann.github.io/
linkedin.com/in/emma-scharfmann/
emma_scharfmann@berkeley.edu

PROFILE

Driven by my interest in AI and innovation, my aim is to improve the accessibility and interpretability of extensive text-based datasets. I am currently working as a Data Scientist at UC Berkeley on research projects in collaboration with Lee Fleming and Matt Marx. I developed end-to-end machine learning models to tackle challenges at the intersection of science (publications), technology (patents) and business. My engineering background and master's degrees in Artificial Intelligence (CentraleSupelec) and Operations Research (UC Berkeley) provide me with the technical expertise to develop models on large unstructured datasets.

EXPERIENCE

- **Data Scientist - UC Berkeley** *August 2022 - August 2024*
  - Three working papers:
    - **Scientist-inventors crosswalk** (first author, submission currently at Science): Created an SQL database of hundreds of millions of papers and patents and developed an end-to-end Random Forest using structured and textual data to identify scientists-inventors, in collaboration with Lee Fleming.
    - **Patent-paper-pairs** (last author, github.com/EmmaScharfmann/PatentPaperPairs): Developed an end-to-end Random Forest using NLP (self-plagiarism and LLM). Data available at relianceonscience.org/patent-paper-pairs.
    - **Image duplicates in Alzheimer's research**: Studied "sloppiness" in Alzheimer's research by analyzing papers' images using ImageTwins software (working paper).
  - Developed an online platform using LLMs to facilitate and optimize the search of scientists, inventors, scientific papers, patents and companies (hosted on HuggingFace: https://emmascharfmannberkeley-synapse-project.hf.space).
  - Supervised 5 groups of Berkeley master's students on machine learning projects (computer vision on patents images, disambiguation of the worldwide patents database using NN with contrastive loss).
  - Collaborated with faculty at UC Berkeley Engineering, Wharton and Cornell Business Schools, and UC Merced Economics.
- **Machine Learning and Optimisation research project** *August 2022 - November 2022*
  - Research project: improved the loss function of Machine Learning models when solving repetitive linear optimisation problems.
- **Teaching assistant in Mathematics - UC Berkeley** *Jan 2022 - May 2022*
  - Taught discussion sections and office hours for 50 undergraduate students (20 hours per week).

EDUCATION

- **Master of Engineering, Operations Research** *August 2022 - May 2022*
  University of California, Berkeley - GPA: 3.91
  - Machine learning; Optimization (linear/non-linear); Applied Stochastic Processes; Mathematical programming.
  - Concentration: Intellectual Property and Entrepreneurship Strategy
- **Master of Engineering, Artificial Intelligence** *Sept 2021 - June 2022*
  Centrale Paris, Paris-Saclay University - GPA: 3.99
  - Statistics and Machine Learning; General Topology; Optimization.
- **Bachelor of Engineering** *Sept 2019 - June 2021*
  Centrale Paris, Paris-Saclay University - GPA: 3.99
  - Functional analysis; Partial Differential Equations; Algorithms.
- **Classe préparatoire, Intensive mathematics and physics program** *Sept 2017 Jun 2019*
  Lycée Henri IV, Paris
  - A two-year post-secondary intensive curriculum in mathematics and physics leading to nationwide competitive entrance examinations to the leading institutions for scientific studies.

ADDITIONAL

- Programming languages: Python, SQL, STATA
- Strong experience with machine learning (random forests, decision trees, gradient boosting) and deep learning (transformers and LLMs, supervised fine-tuning, NN with contrastive loss, CNN) algorithms and tools.
- Experience with GitHub, Google BigQuery/Cloud, HuggingFace, OpenAI API, PostgreSQL, MySQL, multiprocessing.
- Spoken Languages: French (native), English (fluent), German (intermediate).