

# Henri Goua

 Portfolio ·  GitHub ·  LinkedIn ·  gouabeedi@gmail.com ·  Paris, France

## Profil

Diplômé en Data Science (ENSAI), j'ai acquis de bonnes compétences dans la conception de **pipelines Machine Learning et data de bout-en-bout sur le cloud**, avec mise en production d'**APIs ML en Python**. Je possède des connaissances en **statistique, machine learning, deep learning**, ainsi qu'en **NLP, Computer Vision et IA générative**, avec un souci particulier pour l'**évaluation de modèles** et le post-processing. Je recherche une opportunité de **Data Scientist** pour contribuer à des projets d'**IA appliquée et d'industrialisation**, en alliant rigueur analytique et impact métier.

## Formation

**ENSAI — Diplôme d'ingénieur en Data Science et IA** — *bac+5 (Data Science / Ingénieur)* 2022 – 2025  
Spécialisation en **Machine Learning**, NLP et Big Data, avec bases solides en **statistiques, optimisation** et algorithmes.  
**Classes Préparatoires — Mathématiques, Physique & Sciences de l'Ingénieur** 2020 – 2022  
Formation intensive centrée sur la résolution de problèmes et la modélisation.

## Compétences

- **Programmation & Big Data** : Python (pandas, NumPy, SciPy), SQL, *NoSQL* (MongoDB), Spark, Hadoop, R, Git, Bash/Linux ; APIs *REST/ML* ; viz (matplotlib, seaborn, Plotly — *dashboards*)
- **ML/DL & Évaluation** : scikit-learn, PyTorch, TensorFlow, XGBoost ; CNN/LSTM ; explicabilité & évaluation (SHAP, LIME) ; *systèmes de recommandation* (collab./contenu, factorisation)
- **NLP & Génératif** : Hugging Face, spaCy ; LLMs (BERT, T5) ; **RAG** (BM25 + embeddings + reranking) ; LangChain
- **MLOps & Cloud** : FastAPI, Docker, MLflow, Streamlit ; CI/CD (GitHub Actions) ; AWS (EC2), GCP (Cloud Run) ; monitoring (latence p95, qualité)
- **Vecteurs & Recherche** : index vectoriels (FAISS, ChromaDB) ; recherche hybride & reranking
- **Stats, Éco & Séries temporelles** : modèles linéaires/GLM, régularisation, analyse multivariée, économétrie, optimisation convexe, inférence bayésienne ; *ARIMA/SARIMA, VAR, ETS*

## Expérience professionnelle

**Stagiaire — IA Générative & Data Science** Avr 2025 – Présent  
*Square Management — Square Research Center*

- Architecture et **évaluation** d'un pipeline de *document intelligence* (OCR × chunking × embeddings) jusqu'au **déploiement**.
- Conception de **pipelines RAG** de bout en bout (Python) : préparation des données, **retrieval hybride**, génération LLM ; **189+ configurations** testées.
- **Post-traitement** et **recherche hybride** (BM25+embeddings) avec **reranking LLM**, améliorant **robustesse** et **précision**.
- **Évaluation automatique** (RAGAS, TruLens, LLM-as-a-Judge) : **fidélité, pertinence, traçabilité**.
- **Industrialisation** (FastAPI, Docker, CI/CD GitHub Actions) de moteurs vectoriels (ChromaDB, FAISS).
- Livraison d'un **outil de visualisation & curation** avec reporting automatisé (Finance, Marketing, ESG).
- **Documentation** orientée **transparence, auditabilité** et **conformité** (RGPD, biais).

**Stagiaire — Analyse comportementale & Data Science** Juin 2024 – Sep 2024  
*Ville de Paris — Agence de la Mobilité*

- **Exploration et développement** d'un **modèle supervisé** (scikit-learn) pour prédire les flottes de véhicules les plus polluantes.
- Conception d'un **pipeline SQL+Python de Data Processing** pour l'intégration et la correspondance de données (API SIRENE + jeux régionaux), atteignant une **précision de 85% sur 30k+ enregistrements**.
- Mise en place d'un **monitoring (latence p95, robustesse, qualité)** pour assurer la performance et la fiabilité du pipeline.
- Recommandations **data-driven** et **segmentation sectorielle** pour la transition durable et le scoring des émissions.

## Projets

**Moteur de détection de fraude — API de scoring** Jan 2025 – Mars 2025

*Python, scikit-learn, LSTM (séries temporelles), SHAP, FastAPI, Docker, AWS EC2*  
Déploiement d'une API de scoring transactionnel ; AUC-PR = 0,89 sur 285k transactions ; monitoring interactif.

**Modération de contenus — pipeline NLP explicable** Mars 2024 – Mai 2024

*Python, scikit-learn, SHAP, FastAPI, Streamlit, SQL*  
Pipeline de modération explicable ; précision 91% sur 50k avis ; API FastAPI + interface Streamlit.

**Détection d'hallucinations — LLM + évaluation** Déc 2024 – Mars 2025

*HuggingFace (BART/T5), spaCy, Streamlit, OpenAI API*  
RSS→Résumé→Détection d'hallucinations ; 300+ articles ; app d'annotation et workflow reproductible.

**Calibration bayésienne — modèle Lorenz-96** Oct 2024 – Mars 2025

*Python, NumPy, TensorFlow, ABC-Rejection/SMC, MCMC*  
30k simulations ; comparaison des approches ; erreur paramètre  $\approx 0,25$ .

**Vision retail — comptage véhicules satellites** Jan 2024 – Mars 2024

*PyTorch, YOLOv8, OpenCV, GeoPandas, Rasterio, Pandas, LSTM/Prophet*  
Comptage de véhicules sur images satellites (séries de fréquentation) ; nowcasting des ventes ; pipeline (API + dashboard).

## Langues

Français (natif), Anglais (professionnel)