

Baptiste Blouin

Data scientist / NLP, machine learning

Aix-en-Provence, France
✉ BaptisteBlouin@proton.me
🌐 baptisteblouin.fr
in [baptiste-blouin](#)
🔗 [BaptisteBlouin](#)
Français, anglais

Résumé professionnel

Data scientist et docteur en informatique (NLP / machine learning), sept ans d'expérience de la donnée brute au modèle en production. Je construis des jeux de données annotés, j'entraîne et j'adapte des modèles de langue (PyTorch, Hugging Face), y compris sur le supercalculateur Jean-Zay, et je conçois les protocoles d'évaluation qui les jugent. Mes résultats sont publiés : +35 points de F-score face à Jieba en segmentation du chinois, -50 % d'impact des erreurs OCR, 93 % des performances retrouvées avec 250 phrases annotées. Je mets aujourd'hui ces méthodes en production dans Aubier, une plateforme SaaS d'analyse de données par IA (text-to-SQL, RAG, prédictions AutoML).

Projets clés

Dataset NER chinois & pipeline ML

2023–2024

Ingénierie de données & ML

Publié : [LREC-COLING 2024](#)

- **Co-construit le plus grand dataset NER** de presse chinoise historique annoté à la main (1872-1949) : 196 documents, 6 491 entités, liage d'entités et coréférence
- **Développé des systèmes de contrôle qualité** automatisés pour datasets multilingues
- **Créé un pipeline d'entraînement ML** avec méthodologie d'évaluation
- **Stack technique** : Python, frameworks d'annotation, PostgreSQL, pipelines de validation

Modèle de langue chinoise & système de tokenisation

2023

Développement de modèles ML

Publié : [NLP4DH 2023](#)

- **Analysé les motifs linguistiques** du chinois transitionnel par investigation systématique
- **Développé des modèles de tokenisation** atteignant plus de 83 % de F-score, +35 points face à Jieba
- **Collaboré avec les chercheurs** d'Academia Sinica sur ce projet international
- **Stack technique** : PyTorch, TensorFlow, algorithmes de tokenisation, fine-tuning

Simulation d'erreurs OCR & robustesse de modèles

2022

Recherche expérimentale

Publié : [TALN 2022](#)

- **Étudié la performance des modèles ML** sous conditions de données bruitées
- **Développé une approche d'augmentation** réduisant l'impact d'erreur de 50 %
- **Évalué la robustesse des modèles** avec un framework de benchmarking
- **Stack technique** : PyTorch, Transformers, algorithmes de simulation, frameworks d'évaluation

Transfer learning pour l'adaptation de domaine

2021

Développement & optimisation ML

Publié : [NLP4DH 2021](#)

- **Appliqué le few-shot learning** atteignant 93 % de récupération de performance avec données minimales
- **Implémenté des architectures neuronales** character-aware pour traitement de texte en production
- **Évalué les modèles** sur plusieurs datasets avec analyse statistique
- **Stack technique** : BERT, transformers character-aware, techniques d'adaptation

Aubier, plateforme SaaS Data & IA

2025–présent

Conception & développement full-stack

Projet indépendant, aubier-data.fr

- **Plateforme SaaS multi-tenant** de bout en bout : extraction documentaire, RAG, text-to-SQL, analyse et prédictions
- **Moteur RAG avancé** (recherche hybride pgvector + BM25, GraphRAG, RAPTOR) et **text-to-SQL** avec streaming SSE
- **Sécurité complète** : JWT RS256, RBAC, isolation tenant, GDPR, billing Stripe
- **Stack** : FastAPI, PostgreSQL/pgvector, Celery, React 18, TypeScript, LangChain/LangGraph, Docker

HistText : plateforme d'analyse de textes à grande échelle

Recherche & développement full-stack

2023–2025

Publié : JDMDH 2024

- **Plateforme d'analyse de textes** sur 15 millions de documents historiques (20 milliards de mots), 200+ utilisateurs
- **Stack** : Python, React, Rust, R, PostgreSQL, Apache Solr, Docker

Expériences professionnelles

Ingénieur IA & full-stack, Aubier (indépendant) **Sept. 2025 – présent**

- **Ownership produit complet** : de la conception de l'architecture à la mise en production, en autonomie totale
- **Gestion de la complexité technique** : modélisation de données, API, workers async, sécurité, frontend

Data scientist & ingénieur ML, Protisvalor, projet ENP-China, Aix-Marseille Université **Janv. 2023 – août 2025**

- **Responsable technique et développeur principal** de la plateforme HistText, gestion de projet indépendante; contribution à l'obtention d'une bourse ERC Proof of Concept pour sa valorisation
- **Exploration et résolution** de défis ML complexes dans le traitement de textes historiques à grande échelle
- **Communication de concepts techniques** en ateliers et conférences (NLP4DH 2023, LREC-COLING 2024)
- **Collaboration interdisciplinaire** avec historiens, linguistes et informaticiens internationaux
- **Entraînement distribué** de modèles de langue sur le supercalculateur Jean-Zay (IDRIS/GENCI), SLURM

Doctorant & scientifique ML, Protisvalor, LIS / IrAsia, Aix-Marseille Université **Sept. 2019 – déc. 2022**

- **Recherche indépendante** sur subvention ERC avancée européenne, apprentissage auto-dirigé
- **Publié 4 articles à comité de lecture pendant la thèse (7 au total)**, rédaction technique et communication
- **Présenté la recherche** en conférences internationales (PACLIC, NLP4DH, TALN), prise de parole publique
- **Mentoré et collaboré** avec équipes de recherche internationales, leadership interculturel
- **Calcul HPC sur Jean-Zay** (IDRIS/GENCI) : entraînement de modèles NLP (NER, segmentation, transfert)

Compétences

ML & data science	PyTorch, TensorFlow, Hugging Face, scikit-learn, pandas, FLAML / AutoML, Sentence Transformers, NetworkX, UMAP, DuckDB, Analyse statistique
NLP	spaCy, NLTK, NER multilingue, Extraction d'événements, Transfer learning, Adaptation de domaine, Segmentation de mots, Annotation de données, OCR (Docling, Google Cloud Vision)
Données	SQL, PostgreSQL, pgvector, MySQL, MongoDB, Elasticsearch, Apache Solr, Redis
LLM	RAG hybride, GraphRAG, Text-to-SQL, Embeddings, Reranking, LangChain, LangGraph, Agents IA, Évaluation LLM et RAG, Langfuse
Calcul & outils	Jean-Zay / IDRIS (HPC), SLURM, CC-IN2P3, Docker, Linux, Git, CI/CD, FastAPI
Langages	Python, R, SQL, Rust, JavaScript / TypeScript, Java, C++

Formations

Doctorat en informatique (TAL / machine learning), Aix-Marseille Université **2022**

Thèse : Extraction d'événements à partir de fac-similés de documents anciens pour les études en histoire

Compétences de recherche : Investigation indépendante, test d'hypothèses, conception expérimentale, rédaction technique

Directeur : Prof. Benoît Favre | **Projet** : Subvention ERC avancée ENP-China (n° 788476)

Master Informatique (IA/ML), Aix-Marseille Université **2019**

Licence Informatique, Aix-Marseille Université **2016**