

# Chaeyeon SHIM

## Data scientist/Analyst

Spécialisé dans le domaine de l'analyse de données et de l'IA

## Éducation

**Polytech Nice Sophia** 09/2022 - 09/2023

Master 2 Ingénierie - **IA et analyse de données (IA-ID)**

**Université Côte d'Azur** 09/2021 - 07/2022

Master 1 - **Informatique**

**Université Nationale de Kangwon (Corée du Sud)** 03/2014 - 08/2018

Double Licence (bac+4) :

- **Langue et littérature françaises / Convergence Software**

## Projets

### Génération de texte en image/vidéo avec Stable Diffusion

- Generative AI ( Dreambooth , LoRA) Fine-tuning, Prompt engineering

### Documentation Chat Assistant Application

- Langchain , LLM(GPT 3.5 turbo) , VectorDB (Pinecone), Streamlit

### Détection d'image à l'aide de données sur le tissu textile

- Data serialization (TFRecords) , Inception Model, multi-scale feature analysis

## Expériences

**Stage de Fin d'Études - Ingénieur de l'IA / VCLS** 05/2023 - 09/2023

### [Développement du système ACC (Automated Compliance Checking)]

- Développement du système ACC en utilisant la technologie NLP pour améliorer la conformité des rapports de fabrication aux exigences réglementaires dans le domaine de la fabrication pharmaceutique.
- Prétraitement des données, entraînement du modèle sur AWS Sagemaker, Fine-tuning de modèles langage pré-entraînés (BioBert, SciBert).
- Calcul des scores de similarité textuelle sémantique, augmentation des données en utilisant SBERT.

**Stage - NLP Researcher / INRIA** 09/2022 - 03/2023

### [Annotation de textes médicaux] - Python, NER

- Réalisation d'un prototype permettant d'annoter des rapports médicaux du CHU en français.
- Utilisation de dictionnaires (MedDRA) contenant les noms de médicaments et de symptômes, ainsi qu'un réseau neuronal pré-entraîné (SpaCy) et un annotateur SIFR.

**Projet d'Ingénierie des Données en IA** 06/2022 - 09/2022

### [Prévisions de la demande de taxis à New York] - Python, BigQuery

- Data Preprocessing / EDA / Visualization / Feature Engineering
- Améliorer baseline grâce à la modélisation de l'apprentissage machine (Linear model , XGBoost, LightGBM, Random Forest)
- Train/Predict scheduling avec Google Cloud Composer(Airflow)

**Stage - Ingénieur logiciel (Iot)/ Samsung Electronics** 2016 - 2018

### [SCSC - Cours de Convergence des Logiciels Samsung]

- Réalisation d'un programme de formation rigoureux en Convergence des Logiciels (SCSC) chez Samsung Electronics, acquérant une expertise dans divers langages de programmation et outils tels que C, Java, Python et SQL.



☎ (+33) 6 16 27 52 61

✉ chaeyeon2367@gmail.com

🏠 Paris, France

🌐 [www.linkedin.com/in/chaeyeonshim0930](https://www.linkedin.com/in/chaeyeonshim0930)

🔗 <https://github.com/chaeyeon2367>

## Compétences

### [Fonctionnelles]

- Analyse des données
- Visualisation des données
- Machine Learning and Deep Learning
- Text mining
- Natural language processing (NLP) | LLM
- Rédaction de rapport Power BI
- Analyse de séries temporelles
- Model Fine tuning

### [Techniques]

- **Language** : Python | Tensorflow | Pytorch | Langchain | DAX | Excel | Java
- **Data Visualization** : Matplotlib | Ploty | Power BI |
- **DataBase** : MongoDB | OracleSQL | Bigquery | Mysql | VectorDB (Pinecone, Chroma) |
- **DevOps** : GCP | AWS | Sagemaker | Git | Bash |
- **Big data & ETL** : Pyspark | Airflow | Kafka | Shell |

## Langues

- **Coréen** : Natif
- **Français** : Bilingue
- **Anglais** : Courant

## Certifications

 Microsoft Power BI Data Analyst Professional Certificate

 Microsoft certified Data Analysis with Microsoft Excel

 IBM Data Engineering Professional Certificate