

TUNG ANH TRAN

Summary

Data Scientist and applied researcher with experience in statistical modelling, machine learning, and large-scale health and administrative data analysis within national statistics and clinical research institutions. My work has focused on causal inference, longitudinal data analysis, and reproducible data science for public health applications. I am interested in developing statistical and machine learning methods for population health, clinical decision support, and real-world healthcare data.

Research interests

My research interests include causal inference, interrupted time series analysis, implementation science, and machine learning for healthcare. I am particularly interested in developing trustworthy AI methods, including locally fine-tuned large language models (LLMs), to support clinical decision-making, population health, and the analysis of real-world healthcare data.

Skills

- **Programming & Data Science:** Python (Pandas, Numpy, Scikit-learn, PyTorch), PySpark, R (tidyverse), SQL, SAS.
- **Statistical & ML Methods:** Regression modeling, causal inference (Oaxaca–Blinder), classification, time-series forecasting, feature engineering, model evaluation, optimisation.
- **Deep Learning:** CNN architecture design and tuning, image preprocessing, regularisation, hyperparameter tuning.
- **Research Data Management:** Data validation, quality assurance, reproducible workflows, documentation.
- **Cloud & MLOps:** Microsoft Azure (Data Factory, AutoML), Microsoft Fabric (Delta Lake, Dataflows), GCP.
- **Data Visualisation & BI:** RMarkdown, Power BI (DAX, Power Query), Tableau.

Work Experience

DATA SCIENTIST – OUCRU (Oxford University Clinical Research Unit) - VN 2025 - 2026

Responsible for integrated AMS dashboard development and data workflow optimization.

- **Dashboard Engineering & Data Integration:** Developed and maintained the integrated data platform and analytical framework for a 30-page antimicrobial stewardship dashboard used to monitor key stewardship indicators across hospital departments. Integrated four independently developed components into a unified reporting system, ensuring data quality, reproducible workflows, semantic model consistency, and deployment readiness across multiple hospital sites.

DATA SCIENTIST – INSEE (National Institute of Statistics and Economic Studies) - FR 2022 - 2023

Responsible for annual salary database production and research on wage gaps in France.

- **Statistical Modeling and Data Analysis:** Produced and published an official national analysis of public–private wage differentials in France using Oaxaca–Blinder regression on a 40M-record annual salary database, quantifying structural and unexplained components of wage gaps.
- **ETL Pipeline Development:** Designed, implemented, and maintained robust ETL pipelines to standardise occupational data across 1.5M+ records, reducing redundancy by 17% and significantly improving data quality and interoperability across systems.
- **Data Validation:** Investigated and resolved anomalies arising from complex data collection processes, ensuring accuracy, consistency, and reproducibility of official statistics.
- **Reporting:** Analyzed wage dynamics during the COVID-19 period, identifying structural shifts through longitudinal data analysis. Redesigned and presented annual reports on key French wage statistics using RMarkdown.

Education

MASTER OF DATA SCIENCE – Paris 1 Panthéon-Sorbonne University

2023

Majors: Applied Data science in business
Graduated as Valedictorian | GPA: 15.68 / 20
Thesis: Public and Private sector Pay gap in France - Statistical modeling of large administrative datasets | Grade: 17/20

BACHELOR OF ECONOMICS AND MANAGEMENT – Évry Paris-Saclay University 2021
Majors: Economics
Graduated with honor | GPA: 14.93 / 20

Publications

Godet, F. & Tran, T. A. (2025).
For the same annual volume of work, average net wage in the civil service was 3.7% lower than in the private sector. Insee Première, No. 2043, INSEE (French National Institute of Statistics and Economic Studies), 20 March 2025.

Selected Portfolio Projects

- DIGIT RECOGNIZER: MNIST COMPUTER VISION COMPETITION** – Kaggle 2025
- Built and iteratively optimised a CNN model to classify handwritten digits from the MNIST dataset. Applied deep learning techniques including image preprocessing, data augmentation, dropout regularisation, and model evaluation. The final model achieved 99.9% accuracy, ranking in the top 2.5% of the leaderboard as of June 2025.
- ROHLIK: TIME SERIES ORDER FORECASTING** – Personal Project - Kaggle 2024
- Built a custom time series forecasting model using XGBoost with expanding window validation to predict daily warehouse orders, achieving less than 5% MAPE. Engineered time-based and holiday-related features, applied Optuna for hyperparameter tuning to ensure robust performance.
- PREDICTIVE CUSTOMER MODELING** – Personal Project 2024
- Developed a classification model to predict customer response likelihood to a marketing campaign, achieving a 90% ROC-AUC score, and performed precision-recall analysis to evaluate conversion potential. Optimised the final model for profitability, enabling marketing strategies that improved ROI through prediction-driven decision thresholds.

Certifications

- MICROSOFT AZURE DATA SCIENTIST** – Microsoft DP-100 2024
Design, create, manage Azure ML projects; train, monitor, hyper-tune, deploy ML models to batch and online endpoints.
- MICROSOFT POWER BI DATA ANALYST** – Microsoft PL-300 2020
Transform, load data in Power BI; create semantic models and DAX measures; build custom dashboards and reports.

Languages

- **Vietnamese:** Native
- **English:** Fluent (TOEIC 965/990 – C1 level, expired 2024)
- **French:** Professional working proficiency