Jean-François Ton

jeanfrancois287@hotmail.fr

Google Scholar

EDUCATION

UNIVERSITY OF OXFORD | DPhil in Statistics and Machine Learning
Sep 2017 – Sep 2022 | St Peter's College, Oxford, UK

- PhD Thesis title: Causal Reasoning and Meta Learning using Kernel Mean Embeddings
- Supervised by Prof. Yee Whye Teh and Prof. Dino Sejdinovic

UNIVERSITY OF OXFORD | Masters, Applied Statistics
Oct 2016 – Aug 2017 | Somerville College, Oxford, UK

- Distinction (Top 10%) with published thesis in a Journal
- Awarded the Principal's Prize worth £250 (Somerville College)
- Courses include Machine learning, Applied Statistics, Graphical Models, Bayes Methods

IMPERIAL COLLEGE LONDON | Bachelors of Mathematics
Oct 2013 – Aug 2016 | London, UK

- 1st Class Honor ($\sim$ Top 10%)
- Awarded Best First Year Poster Prize in Statistics worth £100 (Score: 100/100)
- Focused on Statistical Methods, project on Markov Chain Monte Carlo Methods

INDUSTRY AND RESEARCH EXPERIENCE

BYTEDANCE RESEARCH/SEED | Senior Research Scientist in Responsible AI for LLMs
April 2022 – Present| London

AMAZON | Research Scientist Intern under Dominik Janzing
Oct 2021 – Feb 2022| Tuebingen

APPLE | Research Scientist Intern Under Joshua Susskind
Jul 2020 – Feb 2021| Cupertino

BLOOMBERG LP | Global Data Intern
Jun 2016 – Sep 2016 | EQUITY EVENTS, London

IMPERIAL COLLEGE | UNDERGRADUATE RESEARCHER
Jun 2016 – Jul 2016 | London
Project: Reinforcement Learning on Games (Supervised by Prof. Calderhead)

IMPERIAL COLLEGE | UNDERGRADUATE RESEARCHER
Jun 2015 – Aug 2015 | London
Project: Creating Unbiased Estimators using Biased Estimators for SDEs (Supervised by Prof. Calderhead)

AWARDS

ESPRC AND MRC STUDENTSHIP FOR DPHIL IN STATISTICS AND MACHINE LEARNING
2017 - 2021 | Oxford, UK

SOMERVILLE COLLEGE PRINCIPAL PRIZE
2017 | Oxford, UK

ARCHIBALD JACKSON PRIZE
2016 - 2017 | Oxford, UK

FIRST YEAR STATISTICS PROJECT PRIZE
2014 | London, UK

PUBLICATIONS ON LARGE LANGUAGE MODELS

ACTIVE REWARD MODELING: ADAPTIVE PREFERENCE LABELING FOR LARGE LANGUAGE MODEL ALIGNMENT | ICML 2025
Yunyi Shen*, Hao Sun*, Jean-Francois Ton

UNDERSTANDING CHAIN-OF-THOUGHT IN LLMS THROUGH INFORMATION THEORY | ICML 2025
Jean-Francois Ton*, Muhammad Faaiz Taufiq, Yang Liu

ACC-DEBATE: AN ACTOR-CRITIC APPROACH TO MULTI-AGENT DEBATE | ICLR 2025
Jean-Francois Ton*, Andrew Estornell*, Yuanshun Yao, Yang Liu

RETHINKING REWARD MODELING IN PREFERENCE-BASED LARGE LANGUAGE MODEL ALIGNMENT | ICLR 2025
Hao Sun*, Yunyi Shen*, Jean-Francois Ton

MITIGATING REWARD OVEROPTIMIZATION VIA LIGHTWEIGHT UNCERTAINTY | NEURIPS 2024
Jean-Francois Ton*, Xiaoying Zhang*, Wei Shen, Hongning Wang, Yang Liu

MEASURING AND REDUCING LLM HALLUCINATION WITHOUT GOLD-STANDARD ANSWERS | ARXIV 2024
Jiaheng Wei, Yuanshun Yao, Jean-Francois Ton, Hongyi Guo, Andrew Estornell, Yang Liu

TRUSTWORTHY LLMS: A SURVEY AND GUIDELINE FOR EVALUATING LARGE LANGUAGE MODELS' ALIGNMENT | ARXIV 2023
Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Hao Cheng, Ruocheng Guo, Yegor Klochkov, Muhammad Faaiz Taufiq, Hang Li

REGULARIZED TRAINING OF NEAREST NEIGHBOR LANGUAGE MODELS | NAACL 2021
Jean-Francois Ton, Walter Talbott, Shuangfei Zhai, Josh Susskind

PUBLICATIONS ON RESPONSIBLE AI

DATASET FAIRNESS: ACHIEVABLE FAIRNESS ON YOUR DATA WITH UTILITY GUARANTEES | NEURIPS 2024
Muhammad Faaiz Taufiq, Jean-Francois Ton, Yang Liu

FAIR CLASSIFIERS THAT ABSTAIN WITHOUT HARM | ICLR 2024
Tongxin Yin, Jean-Francois Ton, Ruocheng Guo, Yuanshun Yao, Mingyan Liu, Yang Liu

FAIR LEARNING TO RANK WITH DISTRIBUTION-FREE RISK CONTROL | ARXIV 2024
Ruocheng Guo, Jean-Francois Ton, Yang Liu

RECTIFYING UNFAIRNESS IN RECOMMENDATION FEEDBACK LOOP | SIGIR 2023
Mengyue Yang, Jun Wang, Jean-Francois Ton

MARGINAL DENSITY RATIO FOR OPE IN CONTEXTUAL BANDITS | NEURIPS 2023
Muhammad Faaiz Taufiq, Arnaud Doucet, Rob Cornish, Jean-Francois Ton

CONFORMAL OFF-POLICY PREDICTION IN CONTEXTUAL BANDITS | NEURIPS 2022
Jean-Francois Ton*, Muhammad Faaiz Taufiq*, Rob Cornish, Yee Whye Teh, Arnaud Doucet

BAYESIMP: UNCERTAINTY FOR CAUSAL DATA FUSION | NEURIPS 2021
Jean-Francois Ton*, Siu Lun Chau*, Javier Gonzalez, Yee Whye Teh, Dino Sejdinovic

SPATIAL MAPPING WITH GAUSSIAN PROCESSES AND NON-STATIONARY FOURIER FEATURES | JOURNAL OF SPATIAL STATISTICS | 2018
Jean-Francois Ton, Seth Flaxman, Dino Sejdinovic, Samir Bhatt

PUBLICATIONS ON META LEARNING

NOISE CONTRASTIVE META-LEARNING FOR CONDITIONAL DENSITY ESTIMATION USING KERNEL MEAN EMBEDDINGS | AISTATS 2021
Jean-Francois Ton, Lucian Chan, Yee Whye Teh, Dino Sejdinovic

META LEARNING FOR CAUSAL DIRECTION | AAAI 2021
Jean-Francois Ton, Dino Sejdinovic, Kenji Fukumizu

METAFUN: META-LEARNING WITH ITERATIVE FUNCTIONAL UPDATES | ICML 2020
Jin Xu, Jean-Francois Ton, Hyunjik Kim, Adam Kosiorek, Yee Whye Teh

PUBLICATIONS ON EFFICIENT LEARNING

GRASSMANN STEIN VARIATIONAL GRADIENT DESCENT | AISTATS 2022
Xing Liu, Harisson Zhu, Jean-Francois Ton, Andrew Duncan

ROBUST PRUNING AT INITIALIZATION | ICLR 2021
Soufiane Hayou, Jean-Francois Ton, Arnaud Doucet, Yee Whye Teh

TOWARDS A UNIFIED ANALYSIS OF RANDOM FOURIER FEATURES | JMLR 2021
Zhu Michael Li, Jean-Francois Ton, Dino Oglic, Dino Sejdinovic

AUTOMATED MODEL SELECTION USING BAYESIAN QUADRATURE | ICML 2019
Henry Chai, Jean-Francois Ton, Roman Garnett, Michael A. Osborne | Long Beach, US

TOWARDS A UNIFIED ANALYSIS OF RANDOM FOURIER FEATURES | ICML 2019
Zhu Michael Li, Jean-Francois Ton, Dino Oglic, Dino Sejdinovic | Long Beach, US

SELECTED EXPERIENCES

TEACHING ADVANCED TOPICS IN STATISTICAL MACHINE LEARNING
Jan 2019 - Jun 2021 | Oxford, UK

- 3x Class tutor for a Machine learning course for fourth year undergraduates.

TREASURER AND CAPTAIN FOR THE OXFORD TABLE TENNIS CLUB
Oct 2018 - 2020 | Oxford, UK

PRESIDENT OF THE IMPERIAL COLLEGE TABLE TENNIS SOCIETY
Mar 2016 - Oct 2016 | London, UK

LANGUAGES / SOFTWARE

PROGRAMMING
Language (in order of experience)
Python • R
Libraries
Pytorch •transformers

SPOKEN & WRITTEN
Native
English
Business
German • French • Luxembourgish