

Thomas GOEURY — PhD Data Scientist

✉ thomas.goeury@protonmail.com • [Linked in](#)



Data scientist et docteur en biologie (génomique des populations) avec de **10 ans d'expérience** dans la conception de pipelines d'analyses et visualisations de données, ainsi que le développement de tableaux de bord interactifs. Passionné par les solutions *data-driven*, j'ai mené à bien plusieurs projets de grande envergure: création d'un **tableau de bord en temps réel** pour le suivi de la pandémie de COVID-19 dans le canton de Genève, implémentation d'un système de **contact-tracing** basé sur Neo4j et développement d'un **algorithme d'analyse de données de séquençage génétique**.

Expérience professionnelle

Université de Genève

Chercheur post-doctoral

Genève (CH)

2023-2025

- Analyses statistique des données de séquençages *next-gen* dans le cadre d'un projet de recherche international étudiant la diversité des gènes HLA (*Human Leukocyte Antigen*) au sein de 40 populations africaines.
- Utilisation d'une méthode inédite basée sur les chaînes de Markov pour analyser l'évolution du complexe majeur d'histocompatibilité (MHC) chez les primates.

Service du medecin Cantonal

Data Scientist, Cellule COVID-19

Genève (CH)

2020-2022

- Mise en place d'un pipeline d'ETL (*Extract, Transform, Load*) et d'un dashboard R-Shiny pour visualiser en temps réel l'évolution de la pandémie de COVID-19 dans le canton de Genève. Amélioration continue de ce dashboard tout au long de la pandémie.
- Développement d'une application de contact-tracing reposant sur une base de données en graphs Neo4j, permettant la détection de nouveaux clusters de contaminations et l'étude des chaînes de contamination.
- Réalisation de rapports et visualisations des données de la pandémie soutenant les décideurs politiques dans leurs décisions (*data-driven decisions*).

Université de Genève

Thèse de doctorat

Genève (CH)

2015-2020

- Développement de l'algorithme MADaM (*Multiplexed Amplicon Data Miner*), pour analyser des données de séquençage *next-gen*, basé sur une classification non supervisée sur des données fortement biaisées; L'algorithme montre une précision de plus de 98.5%, surperformant les solutions existantes de 30%.
- Analyse des gènes HLA sous le prisme de la théorie de l'information (C. Shanon), permettant de cibler plus efficacement les régions géniques les plus informatives, réduisant ainsi les coûts des analyses de génomiques des populations.

Compétences professionnelles

Compétences techniques:

- Langues informatiques: python, R, bash, SQL, neo4j
- Visualisation de données et tableaux de bord interactifs (plotly, shiny, ggplot, dash)
- NLP (*Natural Language Processing*)
- AI: Alg et LLM
- Full-stack *machine learning* et *data science* (sklearn, pandas, data.table, numpy)
- Autres: \LaTeX , Git, environnement UNIX, html/css

Soft skills: Autonomie, travail en équipe, résolution de problèmes, prise de décisions, rédaction de rapports, de synthèses et de documentation

Langues

Français: Langue maternelle

Anglais (professionnel): Lu, parlé et écrit

Autres

2014 - Thèse de master, Grenoble (FR): Modélisation de l'impact du réchauffement climatique sur l'étage alpin des Andes tropicales (páramos) à l'aide des modèles de distribution d'espèce (*SDM*) et de statistiques prédictives

Loisirs: Équitation, maroquinerie (en auto-entrepreneur depuis 2023)