Thomas Charlon, MEng, PhD

Bioinformatics PhD at Quartz Bio, part of Precision For Medicine, and University of Geneva
Master of Engineering in Computer Science at EPITA, France
LinkedIn - **Github** - **Gitlab** - **Google Scholar**

Summary

Experience 10 years of academic and industrial experience in computational data analysis and statistical web applications development

- 1 year as research associate (senior postdoc) at Harvard Medical School CELEHS laboratory (electronic health records analysis, large language models)
- 3 years of entrepreneurship to develop a real-estate market estimation application (geographic information processing systems, self-hosted public-facing web application, nginx, HTTPS)
- 2 years of independent research on withheld content on the Twitter social network (computational linguistics, graph analysis)
- 4 years within the pharmaceutical industry as a Bioinformatician in the Merck Serono spin-off Quartz Bio, now part of Precision For Medicine (genome-wide SNP analysis, unsupervised clustering)

Profile Polyglot expert developer with a passion for reproducible research, aspiring to managerial and leadership roles

- R expert (multiple R packages published on CRAN and cited in peer-reviewed publications)
- Python LLMs / GenAI fluency (Pytorch, Langchain, Cuda)
- Lead strategist and developer of > 10 R/Shiny web applications (Javascript)
- Experience with multiple modern database systems (PostgreSQL, PostGIS, ElasticSearch)
- Server administration (Docker, AWS, cryptography), systems optimization (parallel processing, vectorization)
- Test-driven development, code coverage, reproducible reporting (rmarkdown)
- Experience supervising juniors and other research associates, development of specifications with clients, project management from inception to delivery

Experience

Apr 2024 **Research Associate**; CELEHS laboratory, Harvard Medical School, Boston, MA, USA

Senior postdoc position, 50% on natural language processing of electronic health records for suicide prevention, and 50% on various projects within the lab: developing semantic search prototypes, implementing quality processes, supervising development of statistical visualizations, and creation of standardized codebooks for diagnoses, medications, etc.

- Automated curation of suicide-relevant terms for NLP processing (suicide prevention app to enable clinicians to identify at-risk patients requiring follow-up)
- **Semantic search app** using large language models embeddings (BGE, BERT), comparison of LLMs performances using medical known concepts pairs datasets (PrimeKG) and dimensionality reduction visualizations. Tech stack: Python, Pytorch, Elasticsearch, Nvidia GPU (Cuda), FastAPI, Docker.
- Migration of production Shiny apps to Git source version control and R package structure. Enhancement of visualizations using Javascript. (**Druggable-genome**, **KESER mental health**)
- Creation of 750,000 rows codebook for mapping hospital EHR codes to common ontologies (ICD, Phecodes, RXNORM, CPT, CCS, LOINC), integrated within a **Shiny app** with semantic search capabilities.

Jan - Apr 2024 **Consultant**; Prediction Analytics Research (PARSE Health), Boston, MA, USA

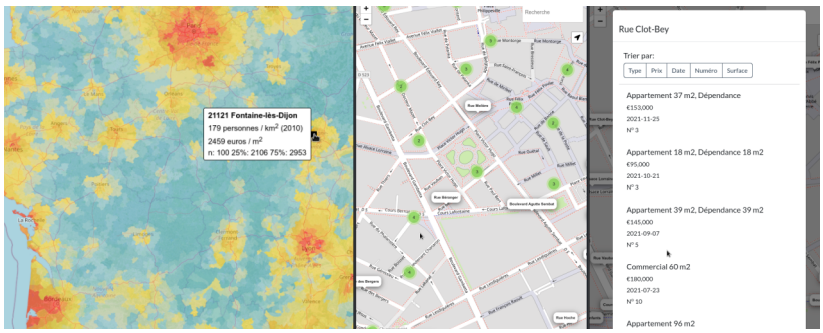
Working with Prof. Tianxi Cai of CELEHS lab (Translational Health Systems) from Harvard Medical School

- Analysis of 8,000 suicide-related open-access peer-reviewed publications (Glove word embeddings, density-based clustering, sparse codings for graphs, novel method development based on random projections to discover clusters of related concepts)
- Setting up implementation quality processes (guidelines for analysis projects development, R packages standards)
- Grant writing (web applications, user-centered design) and Shiny visualization support (large networks, Javascript)

2021 - 2024 **Statistical software entrepreneur**

Real-estate R/Shiny app using open-data (15 million French transactions since 2015)

- Price estimation and market analysis using k-nearest neighbors
- PostGIS database, tailored OpenStreetMap tile server, LeafletJS progressive webapp (PWA)
- Address search using Elasticsearch API (typo correction, location bias)
- Javascript enhancements, visit tracking (Matomo), server administration (nginx, HTTPS)
- Brand identity, business plan analysis, marketing materials, pitch decks



Three screenshots of the Immoservan app

2019 - 2020 **Independent researcher**

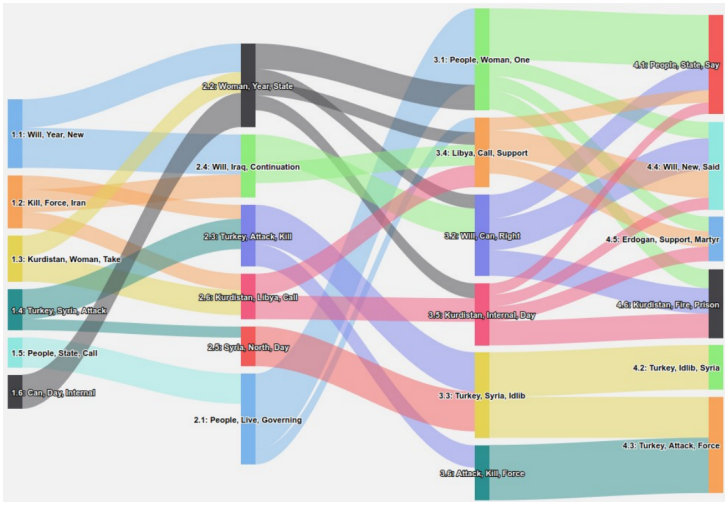
Quantitative trading R/Shiny app

- Statistical arbitrage on cryptocurrencies (Z-scores), mean-reverting strategy

- Automated orders using Poloniex API, backtesting and fine-tuning using Shiny

Monitoring and clustering of Twitter withheld content, using R/Shiny

- Monitoring of 25 European cities during 3 months to discover 79 withheld users
- User graph analysis to discover 2,000 withheld users and 1,400 withheld posts
- Text analysis: Script-based translation, stemming, TF-IDF, LDA topic modeling
- Visualizations: SigmaJS user graph, wordclouds, bi-gram networks, LDAVis
- Novel algorithm to analyze chronological evolution of topics (LDA + Cosine similarity)



Novel algorithm to analyze chronological evolution of topics

2014 - 2018 **Bioinformatician**; Quartz Bio, part of Precision For Medicine, Geneva, Switzerland

Lead genetic analysis of PreciseSADs EU project (collaboration with ELI Lilly, Bayer, and academic institutes)

- Unsupervised clustering of genome-wide SNP microarrays (500,000 markers) of 1,200 cases (RA, SLE, SjS, SSc, Undif.) and controls, characterization of clusters using 100 proteins
- Gaussian mixture models clustering of *HLA* alleles (candidate SNPs)
- Sparse codings using kernel projections and nearest neighbors to increase cluster separability
- Development of modular clustering methods and visualizations to be integrated in internal software base
- Visiting scientist at Genyo institute (Pfizer / University of Granada)

Mar - Aug 2013 **Statistical software developer**; Quartz Bio, Geneva, Switzerland

Development with R of a homogeneous and robust interface to harmonize the statistical frameworks used by the bioinformatics analyses.

Sep - Dec 2011 **Statistical programmer**; Merck Serono, Geneva, Switzerland

Development of a SAS library for CDISC ADaM Analysis Datasets metadata management and submission to US FDA regulatory authority.

Skills

. **Reproducible research, systems and optimization**

- Git code versioning, SQL data processing pipelines, CI/CD, Github Actions
- Server administration, HTTPS, nginx, AWS, uptime monitoring (Netdata)
- Test-driven development, code coverage (testthat, codecov)
- Reproducible reporting and clustering models exploration / backtesting using interactive applications (Rmarkdown, HTML, PDF, Shiny, Javascript)
- Bioinformatics tools: plink, NCBI and BioMart APIs, GWASTools, Bioconductor
- Parallel processing, low-level optimization (SSE/SIMD), vectorization, MapReduce
- Cryptographical tools (eCryptfs filesystems, PGP signatures)

. **Data science and IT side-projects**

- **Personal genome comparison to clinical trials results**
  Compares results of personal genetic tests (*e.g.* 23andme) to a database of 10,000 peer-reviewed genetic variations (SNPedia) and displays matching known genetic variations in a HTML document, sortable by clinical importance, reputation, etc.
- Creation of websites for restaurants and artists. *Go, Javascript*
- Replication of principal component analyses of paintings by **Manovich et al.**.
- **Texas hold 'em poker statistics Android app** *Android SDK, Java*
- Handwriting and speech processing using Markov models. *Matlab*
- Video processing using covariance analysis and mathematical morphology. *C++*

Publications

Peer-review

2017 F1000Research: Replication of the principal component analyses of the human genome diversity panel

Replication of ancestry study using open-data of 500,000 SNPs from 1,000 worldwide controls, available on **Github** with a **Docker image**.

2016 PLOS ONE: Single Nucleotide Polymorphism Clustering in Systemic Autoimmune Diseases

Application of newly developed algorithm to 500,000 genetic variations from 4,100 systemic lupus erythematosus patients and 1,200 healthy controls.

Softwares (CRAN R packages)

2024: Sgraph network visualization

Javascript graph visualizations for large networks of +1,000 nodes. R/Shiny interface to Sigma.JS.

2023: SNPLinkage linkage disequilibrium visualizations

Genetic visualizations combining correlation matrices with chromosomic positions, association studies results, and BioMart gene names. 500 monthly downloads.

2019: OPTICS k-Xi density-based clustering

Specify number of clusters to extract from OPTICS reachability plots, find optimal models and fine-tune parameters using distance-based metrics.

15,000 downloads, used to produce results in proteomics (Nature) and geostatistics (journals, PhD).

Softwares (Github R packages)

2024: Kgraph knowledge graphs and **NLPEmbeds NLP embeddings**

Packages to create knowledge graphs and compute word embeddings (pair-wise word similarities) in unstructured text. Shiny apps for knowledge graphs creation based on cosine similarities or p-value associations.

2017: Mastodon API client

API client for the federated micro-blogging social network Mastodon, 35 stars on Github.

2015: SNPClust, single nucleotide polymorphism unsupervised clustering

Novel algorithm to reduce ancestry bias and enhance disease-relevant signals by unsupervised feature selection and summarization in genetic microarrays based on principal component analysis, Gaussian mixture models, and Markov chain Monte Carlo.

Teaching

2024 **R in Medicine** virtual conference, 1 hour tutorial

Word embeddings in mental heath, from exploration to confirmation, towards multidimensional diagnoses

2017 - 2018 **Free software, Linux, and cryptography**; Geneva and Basel, Switzerland

Organization of lectures and workshops in Geneva. Invited workshop speaker at University of Basel.

Languages

French Mother tongue

English Fluent, lived four years in New Jersey, USA

German Classroom study, intermediate

Education

2014 - 2019 **PhD, Computer Science**; Stochastic Information Processing group, Computer Vision and Multimedia Laboratory, University of Geneva, Switzerland

Genetic clustering for the discovery of a new classification of systemic autoimmune diseases

Supervised by Prof. Sviatoslav Voloshinovskiy (information theory)

2010 - 2013 **Master of Engineering, Computer Science**; EPITA, Le Kremlin-Bicêtre, France

Data science, statistics, and machine learning major (SCIA)

2008 - 2010 **Physics and chemistry preparatory classes**; Lycée Champollion, Grenoble, France

Engineering sciences major