

Yassine YOUSFI

Machine Learning Engineer

PhD Student at Binghamton University

Address : 25 Thorp ST, Apt 6T, Binghamton NY 13905

Mobile : (+ 1) 607 232 3528

email : yousfi1@binghamton.edu

My research focus is on Machine Learning and Deep Learning and their applications in digital media security such as Steganography/Steganalysis or Watermarking. I hold a MSc in Computer Science and Machine Learning from École Centrale de Lille - France.

I believe that technology will not only change the world as we know it, but most importantly, it will help us build a better life for each other.

Education

- 2018+ **Binghamton University – Watson School of Engineering** (PhD) in Machine Learning and Deep Learning and their applications in digital media security
- 2014-2018: **Ecole Centrale de Lille (France)**: one of top 10 multidisciplinary French engineering schools (MSc)
Major: Data Science – Machine Learning, Deep Learning, Information and Decision theory
GPA: 3.7/4
- 2012-2014: **Moulay Youssef – Rabat (Morocco) Preparatory Classes, Mathematics, Physics and Computer science** (Ranked 28/3896 on national final contest)
- 2012: **Yakada High school – Rabat (Morocco)** Baccalaureate – high school diploma in Science (Mathematics speciality), graduated with honours

Work experience

- September 2018+ **Graduate Research Assistant - Digital Data Embedding Laboratory (Watson School of Engineering)** (Binghamton NY)

I am part of a small research group specialized in steganography and steganalysis of digital images, digital image watermarking, forensic analysis of digital images (detection of forgeries), and advanced image processing. Projects are mostly supported by the National Science Foundation (NSF) and the Defense Advanced Research Projects Agency (DARPA). My research focuses mostly on:

- Steganalysis using Deep Convolutional Neural Networks in diverse media sources such as images downloaded from the web, which makes steganalysis challenging because they contain many indeterministic components due to the acquisition conditions and to the high diversity and complexity introduced during development from the raw capture, post-processing, editing, and even sharing.
- Ranked 1st out of all research teams (main contributor from DDE) in the ALASKA steganalysis challenge. We used state-of-the-art deep learning architectures (Residual Neural Networks) as well as recent Deep Learning tricks (Curriculum Learning, output calibration, etc.) to detect images that have been tempered with steganography with very low false alarm rate.

All models have been developed using Tensorflow in python.

- September 2017 – August 2018: **Data Scientist (1-year apprenticeship) – Decathlon International** (Lille France)

I was part of a 10 people data science team (16h/week) at Decathlon focusing on building complex data products. My work focuses on using machine learning techniques to build or improve data products:

- Large scale multistep sales forecasting
Using Machine Learning and Deep Learning techniques to forecast the turnover of each department in all French stores (later extended to Spain, Italy and Belgium), which corresponds to more than 35,000 departments. Forecasting was at week level and performed for the next 8 weeks (multistep forecasting). My contribution was to research different models and architectures and find the best trade-off between accuracy and computation complexity.
All models have been developed using the usual Machine Learning libraries (Scikit-Learn, Keras, xgboost, catboost, etc.) in addition to some field specific libraries such as tsfresh.
The best models were those that used statistics extracted from past sales, weather, holidays, either hand-crafted, or learned using 1d convolutions. These models were surprisingly better than Recurrent Neural Networks which are widely used within the forecasting community.
The developed forecasting system was able to perform better than human forecast, and be used as a building block of several other applications (staff planning, local investments, stock adjustments, etc.)
- Learning attribution models in Omni-channel marketing (web, stores)

Using data driven methods to assign credit for sales to various marketing touchpoints.

My contribution was going beyond linear attribution where each touchpoint in the client journey is given equal credit (or a decaying credit as a function of time), that was used before the project.

We used concepts from coalitional game theory (Shapley value), which is the solution used by Google Analytics' Data-Driven Attribution model, to produce more meaningful and robust attributions.

The algorithms were implemented from scratch using R and shared to marketing leaders using R shiny.

- Link prediction between products on graphs

Given a product, the top co-purchased products were defined using queries and an intractable list of "business rules".

My contribution was to introduce graphs as an effective representation of the co-purchasing phenomena, and use similarity scores (e.g. Adamic-Adar index) to predict the co-purchased product. Co-purchasing graphs also give insights about central products (using centrality scores), finding and harnessing products communities etc. But most importantly, this representation doesn't need any hard coded business rule and is completely learnable from the data.

A visualization tool was also created to help the product developers see how their products interact with each other. The tool was developed using Dash and deployed using Flask and the algorithms were written in the NetworkKit wrappers in Python.

- April 2016 – February 2017 / May – August 2017: **Anti money laundering and counter terrorism financing Consultant (10-month Internship) – Deloitte Advisory & Consulting** (Luxembourg Luxembourg)

I was part of a 30 people team doing consulting activities in fields related to compliance and EU regulations in financial institutions. I specialized in Anti money laundering and counter terrorism financing regulations, and more precisely in achieving compliance through systems and data driven algorithms. My work mainly focused on name matching algorithms;

- Improving name matching performances (entity name screening algorithms against large sanction lists)
Name matching algorithms usually use fuzzy logic to measure the similarity between names. My goal was to reduce the false alarm rate of Deloitte's name screening service (uComply). I used a set of fuzzy If-Then rules based on the Levenshtein distance and other features to reduce the false positive rate with equivalent hit rate
- Crash testing anti money laundering systems of several financial sector companies
In order to assess the compliance of a financial institution's anti money laundering system, I designed a "crash test" scenario on name matching systems. I used a toy names dataset introducing name degradations (e.g. misspelling, inversion, substring, collate, etc.) and monitored the precision/recall accordingly

Skills/Languages

- Programming and development languages/tools related to data science
 - Python, R, Matlab, Spark
 - UML, BPMN, SQL, MongoDB
- Machine Learning – Reinforcement Learning – Deep Learning (Tensorflow, Keras, Pytorch)
- Signal and image processing
- Languages: **French, Arabic:** Native – **English:** Full professional proficiency – **German, Hebrew:** Elementary

Personal interests

- Music: Jazz, Blues. Played instruments: drums, guitar, trombone
- Visual arts: acrylic painting, abstract art
- 2014-2015: member of the Arts Bureau of Centrale Lille

Publications

- Steganography, steganalysis and digital media security

Yousfi Y., Butora J, Fridrich J., Giboulot Q. "[Breaking Alaska Color Separation for Steganalysis in JPEG Domain.](#)" ACM IHMMSEC 2019, Paris.

- Policy and organizational transformation

Martino P., Laban R. K., Yousfi Y., Werbrouck S. "[Organizational transformation in the EU context.](#)" Deloitte Inside magazine – EU special edition.

- Independent articles

Yousfi Y., Bouziane M. "[Optimistic Q-Learning.](#)" Short independent article about Exploration/Exploitation trade-off in Reinforcement Learning.