

# Mario Muñoz Serrano

## Data Scientist

+34 664 633 542 | mariomunozserrano@gmail.com

### WORK EXPERIENCE

#### Business Intelligence Consultant

June 2023 - Now

##### *Holistic Data Solutions*

Develop tools for different clients, to help them understand and optimize their businesses.

- Developed a budget planning web app for an energy provider, offering real-time tracking of departmental budget validation, progress status, and key KPIs such as **EBITDA**. **Board** for the web app.
- **Monte carlo simulations** for optimizing stock portfolio of a small business company. **Python** (pandas, numpy, SciPy and plotly).
- Different **Tableau** dashboards made for a retail company to help them understand and improve their sales distribution on space and time.

**Microsoft Azure** and **SQL** for data ingestion (Connect platform with servers of the client) and python for ETL's.

#### Data Scientist

August 2021 - May 2023

##### *CaixaBank*

Evaluate the economic feasibility of potential fraud prevention strategies and developed fraud detection models for different tasks.

- Detecting and reducing low risk false positives without increase on fraud rates. Using **supervised machine learning** models like Support Vector Machines, Random Forests and Gradient Boosting Machines (e.g., XGBoost, LightGBM).

**Oracle DB** and **SQL** for data modelling, data collection and data quality. **Python** for data processing and deal with imbalanced data (Pandas, numpy), exploratory data analysis (matplotlib, seaborn and bokeh), training models hyperparameter optimization and performance analysis (confusion matrix, precision, recall, ROC-AUC, log-loss) (Sklearn, xgboost, lightgbm and Plotly) to select the best model. And Plotly and shap visualisations to understand and **explain to stakeholders**.

- Detected an extra cost of 2 million € on coin transportation by looking to the **Click Sense** dashboards. To understand causes and reduce them we used **unsupervised machine learning** models like KMeans, Hierarchical Clustering and DBSCAN.

**Microsoft Azure** and **SQL** for data collection and data quality. **Python** for data processing, and data normalization (Pandas, numpy), exploratory data analysis (matplotlib, seaborn and bokeh). Sklearn for PCA dimensionality reduction and KMeans and silhouette score to decide best number of clusters. Scipy for hierarchical clustering by taking a look to the dendrogram of the hierarchical clustering and comparing them with the silhouette score for different k of k-means we have selected the best number of clusters. But used dbSCAN to see if it arises with different conclusions. Plotly to compare the clusters of the three models and some important visualisations like silhouette score visualisation, Within-Cluster Sum of Squares vs number of clusters (elbow method) and dendrogram visualisation.

Related to work methodologies at Fraud & Operations Compliance Team we worked under Agile methodologies (**Scrum**)

#### Data Scientist

November 2022 - January 2023

##### *Boston University*

At the Keck Laboratory for Network Physiology Research Group

Analyse the interaction networks of myoelectrical rhythms across muscles and their evolution with accumulation of fatigue during exercise. Gathered electromyography (EMG) data from various muscle groups during diverse tasks. Used the Discrete Fourier Transform to analyze the **signals**, breaking them down into different frequency bands corresponding to specific muscle fiber types. Computed correlations between these frequency-specific signals, to assess how muscles coordinate during tasks, and how this coordination evolves with fatigue accumulation. **Matlab** for data processing, data transformation, correlations computation and networks visualisations.

### EDUCATION

Mathematical Engineering on Data Science BEng. Pompeu Fabra University 2019-2024

Relevant courses:

#### *Computer Science and Programming*

- Object Oriented Programming
- Algorithm design

#### *Data Infrastructure*

- Networks Architecture
- Databases
- Introduction to Parallel and Distributed Programming
- Cryptography and security

#### *Statistics*

- Probability
- Statistics
- Statistical Models

#### *Machine Learning and Data Science*

- Machine Learning
- Optimization Techniques
- Introduction to Network Science
- Massive Datasets Mining

Course in Artificial Intelligence and Data Science

[Go to the course page](#)

**Artificial Intelligence** text to image, starting from making **API requests** to Midjourney. Ending with the usage of **deep learning** open source huggingface models and stable diffusion models to generate images. Usage of Pytorch to load models, safetensors files and train **LoRa** models. Use of open source datasets, google.colab and imjoy-elfinder a python package that allows you to manage data on remote jupyter servers.

MOOC Elements of AI. University of Helsinki.

[Go to the course page](#)

Political and Administration Sciences BA. University of Barcelona

### LANGUAGES

- Catalan (Native)
- Spanish (Native)
- English (B2)
- French (B1)

## PROJECTS

### NBA Talent Detection Webapp



#### [Go to NBA careers streamlit webapp](#)

Developed a tool to evaluate the success of NBA players' careers. To **collect and build the datasets** for predicting NBA player careers, we used the **nba-api** (**python** package) and pandas. Supplemented with data scraped from Basketball Reference using **beautiful soup**, **json**, and **lxml**. To combine data coming from two different sources requests with **SPARQL** queries to the WikiData query service were used. For data quality we used two approaches rule-based and validation dataset. More details on our [Towards Data Science article](#). Third, for target variable computation (Career Outcome) **tidyverse** and **dplyr** **R** packages were used to first rescale stats in NBA seasons with fewer games played to ensure fair comparisons. And then compute the season outcome and career outcome for each player. [You can see the code here](#)

Finally, three **supervised models** were used (**Random Forest** with **sklearn**, **lightgbm** and **xgboost**) to predict the career outcome of a player. Performance metrics were analysed and the model with better performance (**xgboost**) was stored in a pickle file with all the encoders. **Plotly** and **shap** were used for **explainability** plots. And player images were obtained with request to the NBA cdn and displayed with **pillow**. **Streamlit** cloud and **github** were used for hosting.

### Human Figure Recreation through 3D Graphics Engine

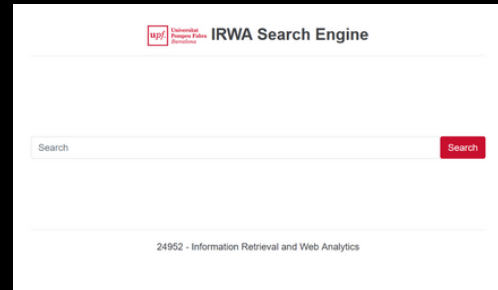


#### [Short demonstration video](#)

#### [Go to the github repository](#)

The project begins with **CPU-rendered** basic shapes, evolving into **GPU-driven 3D scene painting** with a focus on **rasterization**, **sampling**, and **antialiasing**. The project also touches on **virtual reality** considerations, providing a comprehensive understanding of **computer graphics** encompassing rendering, lighting, modeling, animation, and advanced theoretical frameworks. To do that we used the following tools and libraries: **C++**, **SDL** to have keyboard access and move the camera, **OpenGL** to interact with the computer graphics hardware, including the usage of a mesh that defines the object's 3D structure and a **shader** a **GPU** program for determining the visual appearance.

### Search Engine (Information Retrieval and Web Analysis Course)



#### [Go to the github repository](#)

The project involves the creation of a search engine **Flask** web app that extensively incorporates **NLP** techniques to enhance its functionality and performance, integrating an **optimized search algorithm** that considers the intersection of query terms for improved accuracy. The search algorithm includes functions for term preprocessing, **TF-IDF** index creation, and document ranking, with a results page displaying documents in calculated order. A detailed view option for each document is implemented, providing comprehensive information. **Web analytics** are incorporated, collecting data on **HTTP requests**, clicks, sessions, and user context. The analytics dashboard visualizes collected data, offering insights into user behavior, query frequency, and session statistics. The project also addresses algorithmic inefficiencies, optimizing the speed of ranking and index calculation to enhance overall performance.

To do that we used the following tools and libraries: **Python**, **os**, **datetime**, **JSONEncoder**, **random**, **geopy**, **httpagentparser**, **nltk**, **Flask**, **emoji**, **regex**, **uuid**, **json**, **pickle**, **requests**, **Faker**, **stopwords**, **collections.defaultdict**, **array**, **PorterStemmer**, **math**, **numpy**, **collections**, **time**, **pandas**, **seaborn**, **wordcloud**, **counter** and **matplotlib**.

#### [Click Here and discover more projects on my Project Portfolio web page](#)

**Welcome to my Portfolio Page!**

Hi, I am Mario Muñoz Serrano 😊

I'm a **Junior Data Scientist** who starts his journey in the broad field of **Machine Learning and Artificial Intelligence**

I joined **Holistic Data Solutions** as Business Intelligence consultant in August 2023. There I am working on economic planification platforms for different clients.

Previous to Holistic, I've been working at **Calasbank** as Data Scientist as part of the Fraud and Operations Compliance Team. Helping them to create machine learning models to prevent fraud, for over 2 years. At same time during 3 months, I had the opportunity to help the Keck Laboratory for Network Physiology at **Boston University** and learn more about how muscles coordinate during tasks.

I've completed a bachelor's degree in Political and Administration Sciences at **University of Barcelona**. After that I've been curious about Artificial Intelligence and Data Science by doing the course of **Elements of AI** at **University of Helsinki** and to deepen my understanding on that field I decided on 2019 to start my second bachelor's degree Mathematical Engineering on Data Science at **Pompeu Fabra University**.

Where I use different tools like:

**Html**, **css**, **javascript**, **xpath** selectors, **scrapy** and **selenium** for **web scrapping** and **web development**. **Math**, **scipy**, **sklearn** and surprise python libraries to build a **recommendation engine**. Isolation forests for **anomaly detection**. **Apyori** for association rule mining, **jaccard** similarity and **shingling** to find near-duplicates in text data. **Ggplot** and **rayshader** **R** packages to build 2D and 3D visualizations.

Currently working on a **Real-Time Twitter Stream Processing** with **AWS S3**, **Docker**, and **DynamoDB** for the Large Scale and Distributed Systems subject.