Daniel Noumon

LinkedIn: www.linkedin.com/in/daniel-noumon-a06025151
Website: danielnoumon.com
Github: github.com/DanielNoumon
E-mail: daniel_noumon@hotmail.com

Profile

What motivates me in my work is the visible creation of value through data analysis, models, and visualizations. I get a thrill from solving complex data problems and using state-of-the-art AI techniques to do so. On the other hand, I also get a lot of energy from communicating with and involving stakeholders to achieve results, which I consider to be an essential part of data science. Visualizing my findings helps bridge the gap between technical and non-technical expertise.

To become the most versatile data scientist I can be, I am very driven to develop myself in various areas of data science. This is why I'm interested in a wide range of techniques and different industries. I also love discovering different cultures and enjoy traveling to do so.

Scripting

Python, R, SQL, SPARQL, Snowflake, Pydantic, Openai, Azure ai search, Azure document intelligence, Langchain/Langgraph, MLFlow, DSPy, Tensorflow, Keras, SKlearn, SKimage, MCP, prompt engineering, context engineering, RAG, vector databases, OCR, Advanced Chunking, LLM/RAG evaluatie, LangSmith, finetuning models

Visualization

Power BI, Matplotlib, Qlik sense

Software & tooling:

Git, Azure DevOps (CI/CD), Azure Cloud, Github, Gitlab, Infrastructure as code (IaC) - Terraform, Docker, Postman, FastAPI, LLMs, embedding models, rerankers, Weights & Biases

Reporting

LaTeX

Project methods

Scrum/Agile

Data scraping/cleaning

Beautiful Soup, REST API’s

Education

Pre-Master and Master in Data Science & Society (2021 – 2023)
Tilburg University

The Master's in Data Science & Society is a multidisciplinary program focused on applying mathematics, computer science, and statistics to improve business processes.

Master's Research Paper, TiU International Office: "Classifying Earthquake Damage in Nepal: A Comparative Study of Tree-Based Algorithms and Deep Learning."

Software & Tooling:

Python, R, Latex, Power BI, SQL

Bachelor in International Business & Languages (2016 – 2021)
Amsterdam University of Applied Sciences

The Bachelor's in International Business & Languages is both academic and practice-oriented, focusing on improving business processes in environments with international influences. This is accomplished by using methods at the intersection of economics, management, and cultural relations.

Graduation Project for the electric car charger company Wallbox: A qualitative study on which market entry strategy Wallbox should implement to successfully enter the Benelux market.

Work Experience

Data Scientist/AI Engineer (January 2024 – Present)

Data Science Lab, Amsterdam

As a consultant at Data Science Lab, an AI consultancy firm in the Netherlands, I helped lead the Natural Language Processing (NLP) module. This self-initiated team of five consultants specialized in advanced AI techniques for text processing.

Within this module, I developed a elaborate technical foundation for working with Large Language Models (LLMs). I've created a repository where basic techniques such as query transformation, up to complex techniques such as automatic advanced chunking are located in modular fashion for chatbot integration. I also shared my knowledge by organizing and leading a comprehensive NLP training day for the entire technical team at DSL, covering a wide range of topics from basic techniques like tokenization and Word2Vec to the Transformer architecture and topic modeling with BERTopic.

Additionally, I was responsible for six complete proposal processes. In this role, I spoke with prospects about their business problems, determined the technical scope, drafted detailed agile proposals with deliverables per sprint, and presented the proposals. This resulted in converting six out of seven prospects into paying clients, with project values ranging from €25,000 to €103,000.

As a result of the successful growth of the NLP module, our division was promoted in 2025 to an official chapter of Data Science Lab, becoming one of the four pillars of the organization alongside data engineering, data science, and strategy.

Projects at Data Science Lab

AI Engineer (2026 - Present)

Independer, Customer Service Analysis, Hilversum

Independer is a leading Dutch online comparison platform that helps consumers make informed decisions on financial products such as insurance, mortgages, and energy contracts.

I'm building and maintaining scalable AI pipelines to classify and analyze tens of thousands of customer care conversations using speech-to-text systems and large language models.

These pipelines operate on both large-scale historical datasets and daily streaming data, enabling continuous monitoring, insight extraction, and quality analysis.

In parallel, developing automation solutions for other business workflows, including contract document extraction and structured data retrieval, leveraging modern NLP and information extraction techniques to reduce manual processing and improve operational efficiency.

Activities:

- Developing a script that transcribes customer service audio files using open-source software, as well as the Azure Speech Service
- Experimenting with STT / ASR techniques such as diarization to optimize transcription quality
- Experimenting with audio preprocessing techniques
- Implementing statistical analysis on a large scale corpora of transcripts
- Generating dashboards for management and stakeholders
- Constructing Terraform foundation for the Independer GenAI team

Software & tooling:

Git, Python, Terraform, Azure Speech Service, Whisper open source, Pyannote, MLflow, Azure Cloud, LLMs, Pydantic, PowerBI, Microsoft Copilot Studio, Microsoft Power Automate

AI Engineer (2025)
Fokker, HR chatbot, Amsterdam

For aircraft maintenance company Fokker Services Group, I developed a chatbot to support the highly operationally burdened People & Culture department (HR) by handling incoming questions. For this, I was able to use many of the same techniques as for DSL's internal chatbot, with additional methods tailored to the types of documents Fokker wanted to process in the chatbot. This included integrating documents containing many tables, infographics, and images, which introduce extra complexity for RAG applications. To address this, I built customized preprocessing scripts to extract this information as reliably as possible (including OCR & advanced chunking) and convert it into a format that can be accurately and reliably stored in a vector store.

Activities:

- Developing a preprocessing script to extract tables, table images, and infographics embedded in documents
- Developing the RAG model by effectively combining state-of-the-art techniques
- Creating an evaluation dataset based on multiple indicators (information from text, tables, infographics, conversation history, multi-hop questions, guardrails, etc.)
- Structured evaluation with MLFlow to evaluate the retriever (are the correct documents retrieved as context) and the generator (is the correct answer generated based on the retrieved documents)
- Integrating guardrails (against off-topic answers, hallucinations, prompt injection, etc.)
- Memory management
- Logging and monitoring conversation traces for optimization
- Optimizing latency (fast response time)
- Multilingual effectiveness
- Developing a front-end

Software & tooling:
Git, Azure DevOps (Scrum, CI/CD), RAG, MLFlow, Azure Cloud, LLMs, Pydantic, MCP, OCR

AI Engineer (2025)
Fokker, Monitoring Agent/News Summarizer, Amsterdam

For aircraft maintenance company Fokker Services Group, I built a news summarizer that automatically scrapes and summarizes articles from predefined websites. It needed to be used dynamically: by entering a company name (Fokker itself / a competitor / a product name), the user could select different websites, choose a time interval, and select a language. For this, I built a web scraper using BeautifulSoup and a headless browser to retrieve potential web pages. I then filtered these pages for relevance by applying keyword filtering and LLM-based relevancy filtering. Finally, the relevant pages were summarized, key points were extracted, and everything was saved into a document with source attribution. This summarizer could be operated through a web page as the front-end.

Activities:

- Developing a script that dynamically searches for relevant web pages via a headless browser based on 3 parameters (name, websites, time interval)
- Scraping the relevant web pages with BeautifulSoup
- Developing a post-processing script that filters out irrelevant web pages using keyword filtering & LLM filtering
- Generating reports in multiple languages
- Logging and monitoring traces of filtering techniques and report generation for optimization
- Developing a front-end

Software & tooling:
Git, BeautifulSoup, headless browser, MLflow, Azure Cloud, LLMs, Pydantic, MCP

Data Scientist/AI Engineer (2025)

Data Science Lab, Internal Chatbot, Amsterdam

I developed an internal chatbot for operational use within the data science lab. I connected it to SharePoint for document navigation using the Model Context Protocol (MCP) and enabled it to answer questions about document content using Retrieval Augmented Generation (RAG). This allows employees to ask questions about company policy described in HR documents, documentation on completed projects, and onboarding support.

To effectively search through a high volume of documents, I used various retrieval techniques to fetch the correct information. By combining state-of-the-art (SOTA) techniques like hybrid search (vector search + BM25), reranking, reciprocal rank fusion, and metadata filtering, I was able to achieve substantially more consistent and accurate results compared to an out-of-the-box solution like Copilot.

Activities:

- Developing a preprocessing script to add metadata for documents with ambiguous titles
- Developing the RAG model by effectively combining SOTA techniques
- Creating an evaluation dataset
- Structured evaluation with MLFlow to assess the retriever (are the right documents being retrieved as context?) and the generator (is the correct answer being generated based on the retrieved documents?)
- Integrating guardrails (against off-topic answers, hallucinations, prompt injection, etc.)
- Memory management
- Setting up a CI/CD pipeline in Azure
- Deployment via Azure and Docker
- Logging and monitoring conversation traces in production for further optimization
- Developing a secure front-end
- Constructing Terraform foundation for the GenAI chapter of Data science lab

Software & Tools:

Git, Azure DevOps (Scrum, CI/CD), RAG, MLFlow, Azure Cloud, Docker, LLMs, Pydantic, MCP

Data Scientist/AI Engineer (2025)

Data Science Lab, WUA, Amsterdam

WUA helps companies succeed in the online market by tracking and comparing websites. They gather insights into the customer journey through questionnaires given to respondents, which they use to provide strategic advice and practical recommendations. WUA gathers insights into the digital customer journeys of their clients—such as ING, Rabobank, and Vodafone—by conducting surveys with large numbers of respondents.

We created an AI interviewer that replaced the previously static questionnaire with a dynamic interview setting. This solution was developed with a state-of-the-art LLM algorithm, which I specifically optimized for WUA's needs. This allows WUA to collect more relevant and detailed answers, as follow-up questions are asked based on previous responses. This enables them to generate deeper insights into the customer journey and thus create more value for their clients.

Activities:

- Developing a preprocessing script
- Developing the API
- Working with OTAP segmentation
- Setting up a CI/CD pipeline in Azure
- Testing the API
- Scaling API concurrent requests handling
- Analyzing the results with a focus on optimizing metrics in the confusion matrix
- Creating a 'Question answer' LLM model based on industry techniques
- Memory management
- Implementing guardrails

Software & Tools:

Git, Azure DevOps (Scrum, CI/CD), FastAPI, Postman, Azure Cloud, Docker, LLMs

Data Scientist/AI Engineer (2025)

Data Science Lab, Zorginstituut Nederland, Diemen

The "Keteninformatie Kwaliteit Verpleeghuiszorg (KIK-V)" (Chain Information Quality Nursing Home Care) is one of the Zorginstituut's programs. KIK-V's goal is to make reliable and comparable quality information accessible to clients and their families, as well as to professionals. They must answer questions from specific parties who are authorized to request information about quality, finances, etc. KIK-V wants this process to be structured, standardized, and smooth. For this, data is requested from healthcare providers and stored with a central administrator in a data station.

To verify how much of the healthcare provider's data matched the required data in the standardized data station, domain experts from the data station administrator would manually sift through the data to check for a match. This was an inefficient and costly process.

To automate this, I created a matching tool. Using LLMs, data station target data, and healthcare provider input data, the tool checks each required column and its content to see if the data is present in the healthcare provider's input. Additional context was added via a general ontology, where necessary terms are described in detail to support ambiguous column names and increase the accuracy of the matches.

Activities:

- Developing a preprocessing script for the ontologies
- Developing the workflow to match external data to the correct columns via metadata filtering
- Creating a Matching LLM model based on industry techniques
- Analyzing the results with a focus on optimizing metrics in the confusion matrix
- Developing a front-end for the end-users who manage the data station

Software & Tools:

Git, Azure DevOps (Scrum, CI/CD), Azure Cloud, Docker, LLMs, Pydantic

Data Engineer (2025)

Data Science Lab, Zorginstituut Nederland, Diemen

The "Keteninformatie Kwaliteit Verpleeghuiszorg (KIK-V)" (Chain Information Quality Nursing Home Care) is one of the Zorginstituut's programs. KIK-V focuses on making reliable and comparable quality information accessible to clients and their families, as well as to professionals. They must answer questions from specific parties who are authorized to request information about quality, finances, etc. KIK-V wants this process to be structured, standardized, and smooth.

As a semantic web data engineer on the KIK-V program, I focused on applying the ontology to data from providers. For this, I developed SPARQL queries based on functional descriptions and built according to the latest ontology.

I also created a training program with extensive documentation for this project to provide new employees with a guide to efficiently understand all the details around RDF, ontologies, Graph database tooling, and technical and functional descriptions, enabling them to write SPARQL queries autonomously.

Activities:

- Translating data from an existing provider's structure into RDF format as triples, in accordance with the ontology.
- Creating Python tests for the same SPARQL queries to compare their outcomes.
- Developing SPARQL queries to translate functional analyses into technically correct queries that can be run on RDF data according to the ontology.
- Using Protégé to visualize the ontologies.
- Using Graph databases like GraphDB and Fuseki to load the data and write SPARQL queries.

Software & Tools:

Git, Gitlab (Scrum, CI/CD), SPARQL, GraphDB, Fuseki, Azure Cloud, Protégé

Data Scientist/AI Engineer (2024)

Data Science Lab, Athlon, Almere

Athlon is a lease company for corporate and private car leasing. Each month, Athlon receives hundreds of customer reviews after and even during a lease contract, providing feedback on the lease experience. These reviews consist of structured tables with categories, as well as unstructured text fields with various descriptions of complaints.

Until now, Athlon has only performed a surface-level analysis of this data by clustering the structured information. To achieve a deeper understanding, they requested an automated way to analyze the text to link it to specific topics and actions. A company-wide dashboard was created to visualize department-specific statistics, trends, and deeper insights.

This text processing is being implemented using Natural Language Processing (NLP) techniques such as BERT, and Large Language Models (LLMs), which are AI-based techniques specialized in handling large pieces of text. The model with the highest performance was put into production via an API to classify reviews immediately and visualize the results.

Activities:

- Developing a preprocessing script
- Analyzing datasets with a focus on finding correlations and other relationships
- Creating a "Topic Modeling" model based on machine learning
- Training the machine learning model to estimate the topic and subtopics of a review
- Developing the API
- Creating a dashboard to visualize relevant review statistics

Software & Tools:

Git, Azure DevOps (Scrum, CI/CD), Snowflake Python, REST APIs, Azure Cloud, PowerBI, Qlik Sense, LLMs, Pydantic