A black background with white text

Description automatically generated with low confidence

**MSc Data Science**

**Individual Project Module**

**7PAM2002-0206-2024**

**Department of Physics, Astronomy and Mathematics**

**Project title:**

**"Optimizing Query Translation: Understanding Machine Learning Model Sensitivity and Uncertainty in NLP"**

**Student name:** *Venkatesh Sattenapalli*

**Student ID :** *23030670*

**Supervisor:** *Pushp Tiwari*

**Course:** *MSc Data Science with Sandwich*

# Background research

Translating natural language (NL) queries into SQL is a significant challenge in Natural Language Processing (NLP). This enables users to interact with databases without requiring expertise in structured query languages. Traditional NL-to-SQL approaches, such as semantic parsers and heuristic algorithms, often face limitations in handling complex data transformations, making them less effective for real-world applications (Luo et al., 2022).

Deep learning techniques, particularly neural machine translation (NMT) models, have demonstrated remarkable success in various translation tasks. "Deep neural network powered neural machine translation models have made great strides in many machine translation tasks, which suggests that they might be viable for NL2VIS as well" (Luo et al., 2022). While this statement refers to NL2VIS, the underlying principle holds for NL-to-SQL translation, where transformer-based architectures, such as sequence-to-sequence models, can optimize query conversion accuracy.

Moreover, business intelligence (BI) tools play a crucial role in analyzing and presenting data efficiently. These tools assist organizations in making data-driven decisions by facilitating quick data access, sharing, and visualization (Khatuwal & Puri, 2022). BI dashboard tools such as Tableau, Power BI, and Spago BI empower users to create insightful dashboards without extensive technical knowledge, improving decision-making processes (Khatuwal & Puri, 2022). In the context of NL-to-SQL translation, integrating BI tools with query generation models can enhance the accessibility of structured data, allowing non-technical users to retrieve and visualize insights seamlessly.

In the context of NL-to-SQL models, applying variance-based sensitivity analysis can help evaluate the effect of hyperparameters such as learning rate, attention mechanisms, hidden layer size, and dropout rates on translation accuracy. By leveraging insights from power system modeling—where composite models like the Composite Load Model (CMLD) analyze complex interactions between multiple parameters—similar methods can be applied to NLP models to improve generalizability and robustness (Maldonado & Anitescu, 2020). Additionally, modern Bayesian optimization techniques and Monte Carlo methods can be used to assess uncertainty in NLP models, ensuring that they remain stable under various input conditions.

# Existing Approaches

Parameter sensitivity analysis (PSA) is a critical technique in assessing the influence of various parameters on the performance of machine learning models. Sensitivity analysis enables researchers to quantify how changes in specific parameters affect model outcomes, aiding in model optimization and robustness enhancement. An active fault diagnosis (AFD) approach, which incorporates sensitivity analysis, has been used to improve fault detection and isolation by generating an excitation input designed offline to maximize sensitivity for each parameter (Gholami, Schioler, and Bak, 2011). By obtaining maximum sensitivity, precision in parameter estimation is improved, leading to better fault detection.

Understanding the impact of different parameters on query translation accuracy in Natural Language to SQL (NL-to-SQL) models requires a systematic approach to sensitivity analysis. Sensitivity analysis is a powerful technique that helps quantify how parametric variations influence a model's output, offering insights into model robustness and performance (Maldonado & Anitescu, 2020). Traditionally, local derivative-based sensitivity analysis methods have been used in power system research to analyze transient behaviors and model uncertainties. These approaches explore parameter variations around a specific operating point, making them suitable for systems with moderate complexity (Maldonado & Anitescu, 2020). However, global sensitivity analysis techniques, such as variance-based methods, provide deeper insights by studying the effect of the entire parameter space rather than isolated variations.

Uncertainty quantification plays a crucial role in understanding model variability and performance. Recent advancements in stochastic mathematical models and polynomial chaos expansion (PCE) have enabled more efficient uncertainty analysis by reducing computational costs compared to traditional Markov Chain Monte Carlo (MCMC) methods (Takahashi et al., 2024). The variance-based sensitivity analysis using Sobol indices provides an effective way to identify key hyperparameters that significantly impact query translation performance. Applying these techniques to NL-to-SQL translation can lead to more interpretable, robust, and accurate models, helping mitigate biases and optimize performance for real-world applications.

# Research Question

*1. How can parameter sensitivity analysis be applied to machine learning models used in natural language processing to understand the impact of different parameters on query translation accuracy?*

*2. What are the sources of uncertainty in the translation of natural language queries to SQL queries, and how can these uncertainties be quantified using mathematical principles?*

*3. What are the key mathematical principles that can be used to analyze the performance of machine learning models in the context of natural language query translation, and how can these principles be applied to improve model robustness?*

# Project Objectives

My research aims to enhance the accuracy and robustness of Natural Language to SQL (NL-to-SQL) query translation by applying parameter sensitivity analysis and uncertainty quantification techniques. The study leverages variance-based sensitivity analysis and stochastic modeling to evaluate the impact of key hyperparameters on query translation performance and to mitigate uncertainties in predictive models. This will be achieved through the following key objectives:

* + Conduct variance-based sensitivity analysis (e.g., Sobol indices) to measure how key parameters (e.g., learning rate, attention mechanisms, hidden layers) influence SQL query translation accuracy (Maldonado & Anitescu, 2020).
  + Draw insights from global sensitivity analysis techniques used in power systems research and apply them to NLP-based translation models (Maldonado & Anitescu, 2020).
  + Identify sources of uncertainty in query translation, including aleatoric uncertainty (stemming from language variability) and epistemic uncertainty (due to model limitations) (Takahashi et al., 2024).
  + Leverage Polynomial Chaos Expansion (PCE) to model statistical variations in query translation performance efficiently, reducing computational costs compared to traditional Monte Carlo simulations (Takahashi et al., 2024).
  + Utilize probability theory, information theory (entropy and KL divergence), and optimization techniques to enhance model reliability and interpretability (Luo et al., 2022).
  + Implement Bayesian Neural Networks (BNNs) and Monte Carlo Dropout to improve uncertainty-aware training strategies, ensuring more stable query translations (Takahashi et al., 2024).

# Timeline

The following timeline outlines the key phases of the project, including literature review, methodology development, experimentation, and final report preparation.

1. **Literature Review**  
   **Dates**: February 13 – February 25 (13 days)
2. **Methodology Development**  
   **Dates**: February 26 – March 11 (14 days)
3. **Experimentation & Implementation**  
   **Dates**: March 12 – April 5 (25 days)
4. **Preliminary Results Analysis**  
   **Dates**: April 6 – April 13 (7 days)
5. **Final Report Writing**  
   **Dates**: April 13 – April 29 (16 days)
6. **Viva Preparation**  
   **Dates**: May 2 – May 13

# Application to Employment Trends: Dataset Overview

For this analysis, the “Employment and Occupational Classification Dataset” is utilized. This dataset consists of 4,880 records and 7 variables, covering key employment factors such as occupational classification, mode of employment, gender distribution, contract type, nationality, academic year, and the number of employees.

This dataset provides insights into employment patterns across different occupational classifications, helping to analyze trends in workforce composition, contract distribution, and potential disparities in employment opportunities. The data has been pre-processed to ensure consistency and accuracy, facilitating reliable analysis.

# References

Luo, Y., Tang, N., Li, G., Tang, J., Chai, C., & Qin, X. (2022). Natural Language to Visualization by Neural Machine Translation. *IEEE Transactions on Visualization and Computer Graphics, 28*(1), 217-226.

Khatuwal, V. S., & Puri, D. (2022). Business Intelligence Tools for Dashboard Development. *2022 3rd International Conference on Intelligent Engineering and Management (ICIEM)*, London, United Kingdom, 128-131.

Kamath, B. P., Geetha, M., Acharya, U. D., Singh, D., & Rao, A. (2025). Comprehensive Analysis of Word Embedding Models and Design of Effective Feature Vector for Classification of Amazon Product Reviews. *IEEE Access*.

D. A. Maldonado and M. Anitescu, "Variance-Based Sensitivity Analysis of the Composite Dynamic Load Model," *2020 IEEE Power & Energy Society General Meeting (PESGM)*, Montreal, QC, Canada, 2020, pp. 1-5, doi: 10.1109/PESGM41954.2020.9281523.

H. Takahashi, H. Kitagawa, E. Schlaffer, H. Paulitsch and W. Bösch, "Uncertainty Quantification and Variance-Based Sensitivity Analysis for Microstrip Line on IC-Substrate," *2024 International Conference on Broadband Communications for Next Generation Networks and Multimedia Applications (CoBCom)*, Graz, Austria, 2024, pp. 1-7, doi: 10.1109/CoBCom62281.2024.10631259.

M. Gholami, H. Schioler and T. Bak, "Active fault diagnosis for hybrid systems based on sensitivity analysis and EKF," *Proceedings of the 2011 American Control Conference*, San Francisco, CA, USA, 2011, pp. 244-249, doi: 10.1109/ACC.2011.5991038.