


# THÔNG TIN CHUNG CỦA BÁO CÁO

- Link YouTube video của báo cáo: [https://youtu.be/fJ\\_MuK\\_SWR4](https://youtu.be/fJ_MuK_SWR4)
- Link .pdf slides:  
[https://github.com/mausLe/CS2205.MAR2024/blob/main/NL-to-Vis-A\\_Natural\\_Language\\_Interface\\_for\\_Generating\\_Business\\_Data\\_Visualization.pdf](https://github.com/mausLe/CS2205.MAR2024/blob/main/NL-to-Vis-A_Natural_Language_Interface_for_Generating_Business_Data_Visualization.pdf)
- Mỗi thành viên của nhóm điền thông tin vào một dòng theo mẫu bên dưới
- Sau đó điền vào Đề cương nghiên cứu (tối đa 5 trang), rồi chọn Turn in

<ul style="list-style-type: none"><li>• Họ và Tên: Lê Tuấn Anh</li><li>• MSHV: 230104001</li></ul> 	<ul style="list-style-type: none"><li>• Lớp: CS2205.MAR2024</li><li>• Tự đánh giá (điểm tổng kết môn): 9.5/10</li><li>• Số buổi vắng: 0</li><li>• Số câu hỏi QT: 7 (5 Notes + 2 Câu hỏi)</li><li>• Link Github: <a href="https://github.com/mausLe/CS2205.MAR2024">https://github.com/mausLe/CS2205.MAR2024</a></li></ul>
---	---

# RESEARCH PROPOSAL

## TITLE (IN VIETNAMESE)

NL-TO-VIS: TRỰC QUAN HÓA DỮ LIỆU KINH DOANH TỪ TRUY VẤN NGÔN NGỮ TỰ NHIÊN

## TITLE (IN ENGLISH)

NL-TO-VIS: A NATURAL LANGUAGE INTERFACE FOR GENERATING BUSINESS DATA VISUALIZATION

## ABSTRACT

Natural Language to Data Visualization (NL-to-Vis) offers an intuitive approach to data exploration, but current systems struggle with complex user queries. This leads to an under-specified visual generation that does not reflect user information needs. This research addresses this challenge in the context of Business Analytics (BA) by proposing three objectives:

Building the UITxNL-to-Vis Dataset of real-world BA visualizations and queries for NL-to-Vis evaluation and improvement. The dataset and annotations capture user intent and visual specifications. This is a complex and resource-intensive process.

Exploring Advanced NLP Techniques for NL-to-Vis to address limitations in NL-to-Vis, like the NL4DV baseline model. The proposal aims to research the more robust models that can effectively handle complex user queries.

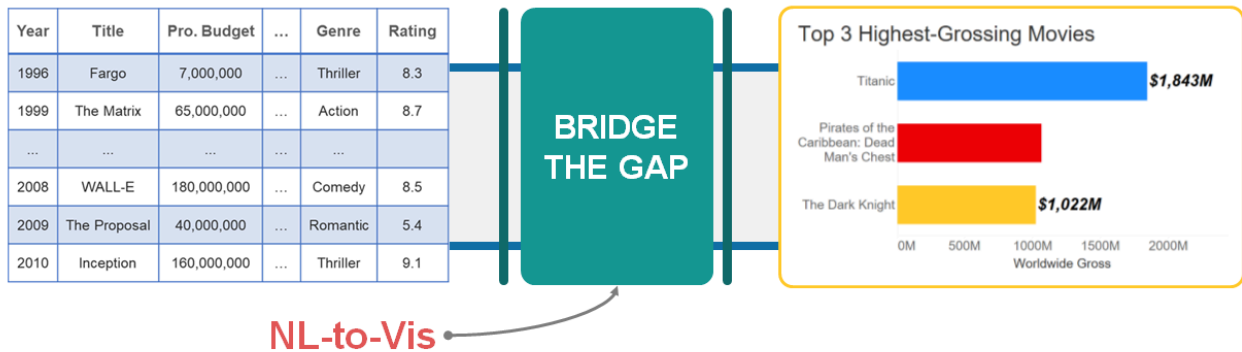
Developing a NL-to-Vis interface, which enables users to explore data through natural language conversation. This makes data exploration more accessible and intuitive regardless of technical expertise in data analytics tools.

## INTRODUCTION

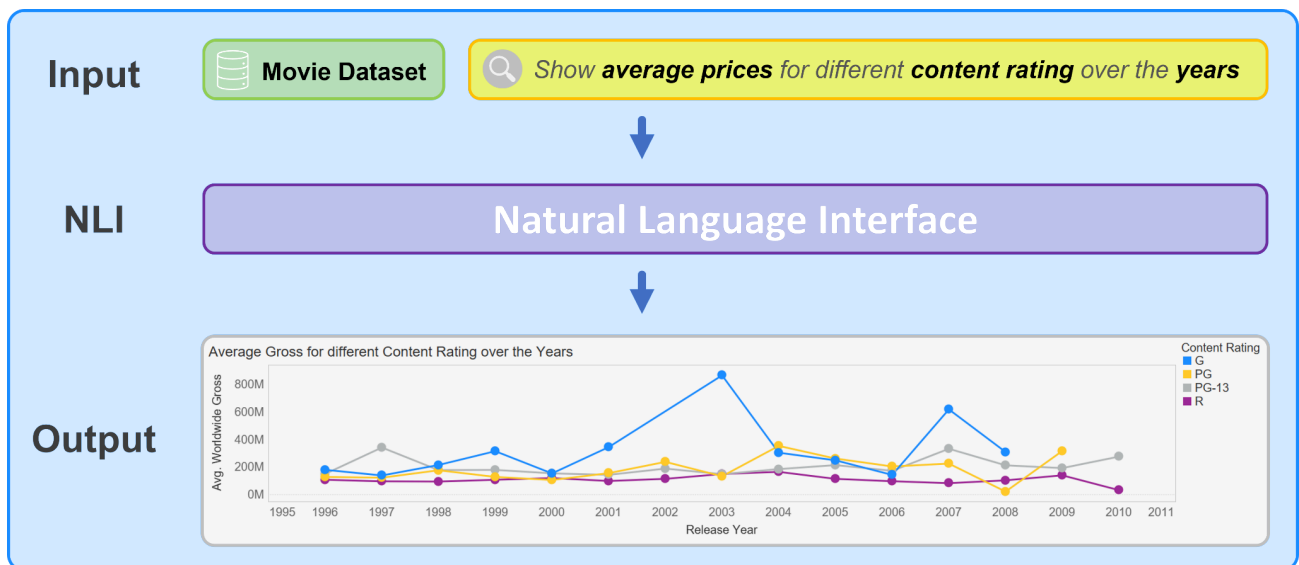
As the world shifts rapidly and data accumulates at high velocity, the ability to quickly analyze and gain valuable insights is crucial for strategic decision-making. However, for many, using data analysis tools to extract insights remains challenging and requires technical expertise. This creates a gap between data and understanding.

Over the past years, Visualization Natural Language Interfaces (NL-to-Vis) have attracted great interest in academia and industry due to their ability to help users flexibly explore, perform analyses using plain language questions, and deliver immediate visual insights without complex data manipulation.

How to Bridge the Gap  
between **DATA** and **INSIGHT**?



A visual is worth a billion data points. This research aims to study NL-to-Vis approaches, which enable users to explore data and gain insights through charts like bar charts, line charts, pie charts... That reflects user information needs.



- Input: There are 2 inputs for an NL-to-Vis:
  - A tabular dataset
  - A text question related to the input dataset
- Output: A visual chart reflects the information needed.

## KEY CHALLENGES

This study is built upon the baseline model NL4DV proposed by Narechania et al. [1] [2]. NL4DV has demonstrated its ability to infer explicit and partially explicit input queries to generate compelling data visualizations.

However, Visualization NLI like NL4DV, XNLI [3], Sevi[4] still faces challenges:

1. **Visualization NLI** struggles with natural language complexity in handling variations and ambiguity in language, requiring contextual understanding.
2. **Under-specification** when NLI needs more information about the aspects of data the user wants to visualize.
3. Evaluating NL-to-Vis is a complex and expensive process due to:
  - **Benchmarking:** Lack of sufficient data to measure how well a system works since collecting user queries and expected visualization is expensive and time-consuming.
  - **User Studies:** Involving human evaluators to provide feedback on usability and identify potential issues is also resource-intensive.

## RESEARCH OBJECTIVES

- **Objective 1.** Building the UITxNL-to-Vis - a Business-Oriented English and Vietnamese Natural Language to Data Visualization dataset. This resource can later be used to evaluate the effectiveness of the research on NL-to-Vis and serve as an evaluation tool for future research.
- **Objective 2.** Using NL4DV [1] [2] as a baseline to explore advanced NLP techniques for effective NL-to-Vis that minimize the impact of ambiguity, variations, and under-specification. These are common challenges for NLI systems like NL4DV.
- **Objective 3.** Developing an NL-to-Vis Interface for data exploration, allowing users to interact with data as having a conversation. This makes data exploration more accessible and intuitive regardless of technical expertise in data analytics tools.

## RESEARCH METHODOLOGIES

- **Objective 1. Building the UITxNL-to-Vis Dataset**
  - Collecting business dashboards from analytics platforms like: Tableau Public, Power BI Gallery using tools like TableauScraper.
  - Data Annotation:
    - Developing a schema to capture user intent and visualization specification (chart type, data type).
    - Manually label the collected data according to the schema.
    - At this step, ChatGPT 4.0 and other large language models could assist in the visual extraction and translation of user intent.
- **Objective 2. Explore Advanced NLP Techniques for NL-to-Vis**
  - Baseline evaluation - NL4DV:
    - Set up NL4DV as the baseline system.
    - Evaluate NL4DV's performance on the UITxNL-to-Vis dataset.
    - Analyze the types of queries where NL4DV struggles, focusing on ambiguity, variations, and under-specification.
  - Exploration of Advanced Techniques
    - Review recent NLP advancements (LLMs & RAG).
    - Consider integration solutions for LLM & RAG to handle ambiguity and under-specification.
- **Objective 3. Develop a Conversational NL-to-Vis Interface**
  - User Interface: Create a natural language data exploration interface. Consider multi-input approaches like: Text input, Voice recognition.
  - User studies could be considered when evaluating the user interface. However, this approach is also a resource-intensive process.

## EXPECTED OUTCOMES

Based on the 3 objectives, the research expects to achieve 3 outcomes:

1. A Valuable Dataset for NL-to-Vis Research

- The UITxNL-to-Vis dataset provides a collection of real-world English and Vietnamese queries paired with corresponding data visualizations from business dashboards.
  - Standardized Annotation with metadata like query intent visualization specification, allowing research to evaluate NL-to-Vis models.
2. Improved NL-to-Vis Capabilities from NL4DV baseline model
    - Minimizing the impact of ambiguity, variations in phrasing, and underspecification in user queries.
    - Outperforming the NL4DV baseline with techniques like LLMs, RAG
  3. A Conversational Data Exploration Interface
    - Allowing users to explore data through natural language conversation.
    - Making data exploration more accessible and intuitive regardless of technical expertise in data analytic tools.

## REFERENCES

- [1]. Arpit Narechania, Arjun Srinivasan, John T. Stasko:  
NL4DV: A Toolkit for Generating Analytic Specifications for Data Visualization from Natural Language Queries. IEEE Trans. Vis. Comput. Graph. 27(2): 369-379 (2021)
- [2]. Rishab Mitra, Arpit Narechania, Alex Endert, John T. Stasko:  
Facilitating Conversational Interaction in Natural Language Interfaces for Visualization. IEEE VIS (Short Papers) 2022: 6-10
- [3]. Yingchaojie Feng, Xingbo Wang, Bo Pan, Kamkwai Wong, Yi Ren, Shi Liu, Zihan Yan, Yuxin Ma, Huamin Qu, Wei Chen:  
XNLI: Explaining and Diagnosing NLI-based Visual Data Analysis. CoRR abs/2301.10385 (2023)
- [4]. Jiawei Tang, Yuyu Luo, Mourad Ouzzani, Guoliang Li, Hongyang Chen:  
Sevi: Speech-to-Visualization through Neural Machine Translation. SIGMOD Conference 2022: 2353-2356