Project Requirements for ENGO645/537 (Winter 2025)

Overview:

The project is a vehicle for the student to take an active part in exploring the subject area, as appropriate for the data mining and spatial databases course. Thus, a project must be chosen that will clearly extend your knowledge and understanding of the data mining and spatial databases area relevant to the course. Moreover, the primary criterion for evaluating your project will be what you have learned and discovered from the project.

There are three aspects to a project:

- Problem definition, literature and technology background.
- Algorithm applied and designed
- System implementation and experimentation

The **problem definition**, literature and technology background should explore and describe the problem or area of interest and provide some investigation of background work in this area. It should examine the research literature and available technologies in the selected area. Because of the time limits, this stage is generally expected to be very limited, e.g. google around, read a couple of papers, study from online tutorial and books, think hard for a while and write your project proposal.

The **algorithm** should try to map out an algorithmic solution that is implementable with the available time and computer resources. With the intended "prototyping" nature of the projects, the design, implementation and evaluation may be performed iteratively.

In the **implementation and experimentation** phase, the algorithm design and implementation need to demonstrate the algorithm feasibility, effectiveness and efficiency.

In general, the course project must be relevant to the issues and techniques and directions covered in the course. But you are allowed to use more advanced data mining, AI and machine learning methods. Original thoughts, solutions, insights are strongly encouraged and will be rewarded.

All work submitted should be done in a professional manner. Documents should be type-written and use good technical English. Software should be well-written, meeting reasonable standards of programming.

Data requirement:

Real datasets are required for the project. The following are some open data websites, but feel free to search and use datasets from other websites. Please remember to review the terms of use for each dataset to ensure compliance with any licensing or usage restrictions. Below is a curated list of repositories for spatial and temporal data:

- UCI Data Repository: A collection of databases, domain theories, and data generators that are used by the machine learning community for the empirical analysis of machine learning algorithms. (https://archive.ics.uci.edu/datasets)
- Data.gov: The U.S. government's open data portal, providing access to thousands of datasets. (https://www.data.gov/)
- European Data Portal: A platform that provides access to open data published by European countries. (https://www.europeandataportal.eu/)
- UN Data: A gateway to the statistical databases of the United Nations. (http://data.un.org/)
- World Bank Open Data: Access to a comprehensive set of international development data. (https://data.worldbank.org/)
- Google Dataset Search: A tool that enables the discovery of datasets stored across the web. (https://datasetsearch.research.google.com/)
- Kaggle Datasets: Kaggle is a platform for predictive modeling and analytics competitions. It also hosts various datasets for exploration. (https://www.kaggle.com/datasets)
- GitHub - Awesome Public Datasets: A very good curated list of high-quality datasets on GitHub. (https://github.com/awesomedata/awesome-public-datasets)
- Open Data Network: A platform that provides access to open datasets from various sources. (https://www.opendatanetwork.com/)
- City of Calgary Open Data: Geospatial and urban data for Calgary, including transportation and zoning. (https://data.calgary.ca/)
- Geolife GPS Trajectory Dataset: GPS trajectory data for spatial and movement analysis.(https://www.microsoft.com/en-us/research/publication/geolife-gps-trajectory-dataset-user-guide/)

Team Formation

To ensure fairness and effective collaboration among students of varying academic levels, team formation will follow these guidelines:

1. **Team Size**:
    - **Undergraduate Students**: Teams should consist of **3–5 members**.
    - **MEng (Course-based Master's)**: Teams should consist of **3–4 members**.
    - **MSc (Thesis-based Master's)**: Teams should consist of **2–3 members**.
    - **PhD Students**: Teams should consist of **1–2 members**.

2. **Mixed-Level Teams (Optional):**
    - Teams may consist of students from multiple academic levels, but the maximum team size must adhere to the higher-level limits (e.g., if a team includes MSc and MEng students, it should not exceed 3 members).
    - **Expectation**: Mixed teams must demonstrate balanced contributions from all members, clearly articulated in the final project deliverables.

3. **Roles and Responsibilities:**
    - Each team member must take on a specific role (such as data preprocessing, algorithm design, implementation, etc.) and these roles must be clearly documented in the project proposal and final report.

- Contributions will be evaluated individually to ensure fairness, so active participation is essential.
- Each team member should contribute equally in terms of time and effort, relative to their academic level.
- Higher-level students (e.g., MSc and PhD) are expected to tackle more complex and in-depth aspects of the project.

4. **Team Collaboration Tools:**
   - Teams are encouraged to use tools like **GitHub**, **Google Drive**, or **Slack** for effective collaboration and version control.

Important Dates:

The project development must be completed by the end of the term, and the following milestones must be adhered to. Each team is required to submit the relevant materials and documents to D2L by the specified dates.

1. Team Registration and Topic Submission (Due Date: February 13, 2025)
Submit a **0.5-page PDF** with team member names, academic levels, email addresses, project title, brief description, and tentative roles.

2. Project Proposal (Due Date: March 10, 2025)
A **1–2 page PDF** including:

- **Problem Statement**: Clear description and relevance.
- **Objectives**: Goals and innovations.
- **Methods**: Spatial data mining techniques and algorithms.
- **Data**: Sources and preprocessing plan.
- **Work Plan**: Timeline and milestones.

3. Progress Presentations (Dates: April 3, April 8, April 10, 2025)
A **10–12 min in-class presentation** with:

- Problem recap, progress, challenges, and next steps.

4. Final Deliverables (Due Date: April 17, 2025 (9:00 AM MST))
A. Presentation Video

- **Duration**: 7–10 minutes.
- Include: Problem, methods, results, insights, and future work.

B. Final Report

- **Length**: 5–10 pages (PDF).
- Sections: Abstract, Introduction, Background, Methods, Implementation, Results, Conclusion, References. Use visuals for clarity (e.g., charts, maps).

Evaluation:

Category	Weight	Description
Innovation & Originality	20%	The project's novelty, creativity in problem-solving, and use of innovative spatial data mining techniques.
Technical Depth	20%	The appropriateness and sophistication of the chosen methods, algorithms, and analysis techniques.
Implementation Quality	20%	The functionality, efficiency, and robustness of the implemented solution, including code quality.
Results & Analysis	20%	Depth of insights derived from the results, supported by meaningful visualizations (e.g., charts, maps).
Presentation & Report Quality	20%	Clarity, professionalism, and organization of the report and video, including proper use of references.

Suggested Project Topics:

The following topics provide **general directions** aligned with the course focus on spatial data mining and databases. These are meant to **inspire** innovative ideas and highlight recent trends in the field. You are encouraged to build on these topics, **refine them to suit your interests**, or **propose your own unique project**. Each topic serves as a starting point—detailing the problem, methods, and expected outcomes will require your critical thinking and creativity.

- **Crime Hotspot Analysis with Predictive Modeling**: Identify urban crime hotspots using clustering techniques and predict future crime-prone areas based on socioeconomic factors and time-series analysis.
- **Delivery Route Optimization Using Trajectory Data**: Analyze delivery trajectories to optimize routes, incorporating travel time and energy efficiency metrics.
- **Detection of Anomalous Traffic Patterns**: Detect unusual patterns in traffic flow using clustering and classification techniques, and correlate anomalies with specific events like accidents or holidays.
- **Spatial Prediction of Air Quality Trends**: Use spatial and temporal data to predict air quality in urban regions, integrating meteorological and pollution sensor data for forecasting.
- **Wildfire Spread Simulation Using Spatial Data**: Model wildfire spread using spatial and temporal data, integrating wind, topography, and vegetation data with machine learning techniques.
- **Wildlife Risk Mapping and Evacuation Planning**: Analyze GPS-tracked wildlife movement to predict future trajectories and plan for evacuation based on the people movement and wildfire spread.
- **Flood Risk Mapping and Evacuation Route Optimization**: Use elevation, rainfall, and river flow data to map flood risks and propose evacuation routes optimized for safety and speed.
- **Optimal Placement of EV Charging Stations**: Develop geospatial models to identify optimal locations for electric vehicle charging stations based on demand forecasting and traffic data.
- **Urban Heat Island Mapping and Mitigation**: Analyze spatial data to identify urban heat islands and evaluate the effectiveness of proposed mitigation strategies, such as green spaces and reflective surfaces.
- **Optimizing Emergency Response Times**: Design a model to minimize emergency response times by optimizing the spatial placement of fire stations, hospitals, or ambulances using real-world data.
- **Public Transit Inefficiency and Route Redesign**: Analyze public transit GPS trajectories to identify inefficiencies, redesign routes, and simulate improvements for better service coverage.

- **Healthcare Accessibility Evaluation Using GIS**: Assess healthcare accessibility by combining population density, transportation networks, and healthcare facility locations, and identify underserved areas.
- **Urban Expansion and Resource Impact Analysis**: Analyze satellite imagery to map urban growth patterns over time and evaluate their impacts on critical resources like water and electricity.
- **Tourism Hotspot Prediction Using Social Media Data**: Analyze geotagged social media posts to identify tourist hotspots and recommend optimal travel routes or itineraries.
- **Mapping Renewable Energy Potential Using Spatial Data**: Combine solar, wind, and land-use data to evaluate and visualize regions most suitable for renewable energy projects.
- **Spatial Noise Pollution Analysis and Mitigation**: Analyze spatial patterns of noise pollution in urban areas and evaluate the effectiveness of potential mitigation strategies like noise barriers or zoning changes.
- **Spatial Modeling of Disease Spread**: Predict the spatial spread of diseases (e.g., COVID-19) using mobility data, population density, and healthcare capacity as key variables.
- **Smart City Planning Using Multi-Layered Spatial Data**: Propose sustainable urban planning solutions by integrating spatial datasets like infrastructure, population, and environmental data for smart city development.
- **Multi-Source Land Use Classification**: Perform land-use classification by integrating satellite imagery, demographic data, and environmental attributes to evaluate urbanization patterns and impacts.

Good Luck!

ENGO 645/537 Teaching Team