Course Project Proposal: Data Visualization Application

<div align="center">February 15, 2025</div>

Steam Games Data Visualization

by Denis Troegubov, Diana Tsoi, Victor Mazanov

Project Goal and Vision:

This project will create a website visualization of dynamics in the gaming industry. We plan to analyze the distribution of genres, what features make some games stand out by creating insightful visualizations, that would also cover how games led to evolution in related domains of knowledge, such as graphics rendering or computer vision or augmented reality. The target audience includes enthusiasts from gaming, as well as researchers in technologies and decision-makers in the industry. The main objective is to provide context for the emerging trends in gaming, that connects it to the business domain.

Dataset Description:

- **Data Source and Collection Plan:**
  - **Web Scraping:** Data will be scraped from Steam website, different monitoring and tech forums, such as pcgamingwiki.com and SteamDB. We will filter data in web-pages to get tables with technical specs and parameters that we need.

- **Data Content:** The scraped data will include:
  - Name

- Price
- Game genre
- User rating
- Number of reviews
- Technical specs
- Release date
- Maximum size of the discount
- Stack of technologies

Visualization App Architecture:

The application will use the standard data visualisation pipeline:

1. **Data Scraping:** Scrapy for web data extraction from Steam and other websites.

2. **Data Cleaning and Preprocessing:** Pandas for handling missing data, inconsistent datatypes and structuring datasets.

3. **Data Exploration and Processing:** Plotly, Seaborn, Matplotlib will be used for the exploratory data analysis to understand price dependencies in different game genres, analyze correlation of popularity, discounts and technical specs, find outliers, aggregating data.

4. **Data Delivery:** Flask RESTful API serving data in JSON format.

5. **Data Visualization:** We'll exploit visualization tools suited for the task - D3.js and P5.js, with Three.js to optionally show the complex relations.

   The main graphical elements we plan to create:

   - Sankey diagram/Tree map: how various genres evolved over time, possibly broken down by the release year of games or the number of active players.
   - Bubble chart with axes representing key game features (e.g., graphics, AI complexity, multiplayer options) and the size of the bubbles indicating success metrics like user ratings or sales.
   - Flow chart: innovations in gaming to advances in related domains (AI, graphics rendering, etc.)

- Optional Panel: Three.js and P5.js (potentially sigma.js) - 3D Network View

Proposed Visualizations and Features:

- **Interactive Interface of the diagrams:** A Sankey diagram with lines of width corresponding to the impact of genres in industry over time. Users can filter by genre, year category, and explore tooltips with numeric data.

- **Automatic flow chart:** User is able to see dynamically evolving technologies in autoplay mode.

- **Possible augmentation with Interactive Network Graph:** Users can click and view the granular information on the impact of nodes to the whole picture.

Project Timeline and Milestones:

- **Week 1-3:** Dataset collection and preprocessing.

- **Week 4-6:** Flask API development and exploratory data analysis.

- **Week 7-9:** Implementation of core visualizations in D3.js, Three.js, Sigma.js.

- **Week 10-11:** Interactive feature development and integration.

- **Week 12:** Final testing, polishing, and presentation preparation.