Group Assignment 1 (G1): Project Characterizations

Interactive-Visual Data Analysis
Fall 2025

DUE: November 9, 23:59

General Task Description

You are now transitioning to the group project phase of the IVDA course! Here, you will form groups of 3 students, with whom you'll work the rest of the term to create your very own IVDA tool.

The course project allows you to apply all the lecture and exercise material to a real-life problem. Through the group assignments and future exercises, you will be guided through building an IVDA tool from scratch. The work you have already done in A1, the quizzes, and the engagement with methods from class will also be quite helpful in this effort. Our goal is to allow each group to work on a problem they deem interesting and inspiring. The IVDA course team really hopes that through this inspiration and work, each of you gets to develop a powerful skill set in IVDA development.

You will start with G1 and characterize the problem space of your topic, using the **what-why-how** method (Nested Model) presented in class.

Point Distribution

This assignment is worth 10 points in total for each student. All answers should be related to your group's topic, and the characterization and abstraction of the domain situation, data, and tasks. Generally, the points are allocated as follows, with a more detailed breakdown in the Written Response Instructions below:

- **Domain:** Characterization of your project (0.5 point)
- **What:** Abstraction of your project's data (3 points)
- **Why:** Abstraction of your project's tasks (3 points)
- **How:** Sketch of the interactive-visual design prototype (3 points)
- **Group Dynamics:** Describe the roles each of your group members will fulfill (0.5 points)
- **References and in-text citation (in written response, don't forget)** (mandatory)
- **Declaration of AI use (in written response, don't forget)** (mandatory)

Submission

Your written response should include:

- A group name – be creative
- The names of all of your group members
- Your written responses to the prompts below (in full sentences, using the *.tex* template)
- Proper referencing of all of your sources used in your written response, excluding lecture and exercise material

You will submit 1 PDF file to your group's OLAT directory, and be graded as a group (i.e., 1 PDF will be provided that represents all 3 of you). Your OLAT group directory will become available to you by October 23, 23:59.

Please submit your written work **on OLAT as a single PDF, using the answer_sheet_G1.tex template** provided for you on OLAT. Please also ensure the text length of your written response is **no more than 4 pages in length**, excluding references or figures. Late submissions face the usual late policy, as outlined in the course syllabus, and will apply to all group members. The proper declaration of AI use by all group members (no use is also an answer) is grading-relevant. Submissions without proper references/citations (APA or IEEE-style) will not be graded. The template files provided to you include an example on including an IEEE-style reference, and more guidance can be found here: `https://www.bibtex.com/s/bibliography-style-ieeetran-ieeetrann/`

Written Response Instructions

In your written response, please describe your approach by answering the following questions. Your submitted response to these prompts should be **no more than 250 words per section**. Please use the provided template to prepare your submission. Include any references that you used; don't forget the declaration of AI use, as discussed in class.

Domain – 0.5P

Briefly describe your project's domain situation, the overall analysis goal, and the domain problem. You should be referring to the information in your topic's description (1 point).

What – 3P

Based on the familiarization with your dataset:

1. Characterize the dataset type of the data source you use in your project. Describe what the data objects are that your analysis will focus on, according to Munzner's VAD Chapter 2 and L02 (0.5 point)
2. Characterize the attributes of the data source you intend to use in your project, according to Munzner's VAD Chapter 2 (1 point)
3. Describe the quality of your data using some summary statistics on the attributes of the data source you intend to use in your project. Include a brief statement on the pre-processing and wrangling steps you performed (or will need to perform) to make the data more usable and useful (more about this in L11, on Oct 23rd), especially in how they relate to your answers in the **How** section of your response below (1.5 points)

Why – 3P

1. Describe your project's target user(s). You should be referring to the information in your topic's problem statement and overall analysis goal (0.5 points)
2. Characterize the core analysis tasks you believe your project is meant to support, according to Munzner's VAD Chapter 3 and L03. To describe tasks, use the tuple notation as presented in the lecture, consisting of an action and a target. Typically, the projects should include around 5 tasks (1 point)
3. For each task you've identified, briefly describe why your group feels it's necessary to support them, in the context of your project's domain and target users. Each task description should include information on the specific data attributes required for its support. For each task, discuss if machine learning (ML) support will be needed to support this task algorithmically (more about ML in L12 and L13) (1.5 points)

How – 3P

Do sketches either by hand or using Figma (or similar):

1. Sketch out (in figures) and describe (in text) your proposed visualizations for supporting the individual tasks identified in the previous step, commenting on possible visual encoding choices according to Munzner's VAD Chapters 5-6, as well as lecture material (L04-L06). Use the **(What + Why) = How** structure of the Nested Model discussed in the lecture to justify your choices on marks, channels, and chart design (1.0 point).
2. Sketch out (in figures) and describe (in text) the combinations of your charts/views into a multi-view IVDA interface (your ideated IVDA solution). Highlight the visual encoding that will link the same information across views, e.g., by using color (0.5 points).
3. Sketch out (in figures) and describe (in text) the interactions that will support your identified tasks and user actions. Discuss at least one interaction that not only directly affects the data representation in the frontend, but also has an effect on the data state or the ML modeling in your backend, e.g., execute a clustering (1 point).

4. Describe (briefly!) your proposed algorithmic modeling/ML method for supporting the tasks identified in the previous step, with a link to a helpful resource or tutorial on the topic (ex., *scikit-learn* documentation). Note that in the IVDA class, you do not need to implement an ML method (too time-intensive). Though later in the group project, you need to demonstrate the use of at least one ML method from an existing library in your IVDA solution (0.5 points).

Group Dynamics – 0.5P

Using the Data Baton definitions and awareness gained in the exercise of (L10), briefly describe the roles each of your group members will fulfill, according to your various skill sets and interests with respect to the project (0.5 points).

References and in-text citation, in written response (mandatory)

Declaration of AI use, in written response, (mandatory)

Project Topic Descriptions - Full List

Your group project in the IVDA course will be based on one of these topics. Below is a full list of the topics available. You will be given a form to fill out to declare your group's choice. Choices are inspired by first-come, first-served, so please pick out at least 3 choices that your group would be happy to work on. We plan to distribute your topic assignments no later than October 23rd, 2025, 17:30. The problem statement you are assigned will also be the topic that you'll characterize in your G1 submission.

The IVDA datasets linked below that refer to Seafile can also be accessed via:

`https://seafile.ifi.uzh.ch/d/5703f1183f2a44beb57d/`

Topic Presentation Schema

- **Topic ID:** XY
- **IVDA Research Stream:** Analytical focus or methodological family if IVDA the project belongs to (e.g., interactive item ranking, exploratory data analysis, relation discovery
- **Domain:** Typically includes the application/domain context of the project, such as finance, healthcare, or sustainability, describing where and why the analysis problem matters.
- **Dataset:** The source and content of the data used (e.g., Airbnb listings, Olympic results, university metrics), outlining what entities are represented and what they describe.
- **Dataset Type:** Includes a rough overview of the data's structural nature (e.g., multivariate, temporal, textual, hierarchical) and determines what analytical and visual techniques are suitable.
- **User group:** The intended users of your IVDA tool (e.g., investors, doctors, students, policymakers) and their expertise or decision-making needs.
- **Motivation/Goal:** Typically includes the problem to solve or insight to enable, such as helping users rank, compare, explore, or explain data interactively.

LLM-Based Data Enrichment Special

Given the growing role of large language models (LLMs) in data science, you may choose from several topics that incorporate an *LLM-Based Data Enrichment* component. This means: an LLM is prompted through an API to provide values for an additional attribute that you aim for (or several). Example: "I also want to have an attribute called *Sustainability Development*" in my IVDA solution about company data. If you plan to apply such an LLM-based enrichment as a preprocessing step (as part of YOUR data abstraction rather than your USERS' activity during the use of your IVDA solution), we offer a streamlined workflow in collaboration with an experienced IVDA team member and using our internal resources. When enriching your dataset with additional attributes through this process, the entire procedure is expected to take you less than one additional hour, at no monetary cost. Just reach out to us, hand in your targeted data, express your attribute wishes, and receive an enriched dataset back (after some compute delay).

Item Labeling: Apartment Ratings in Zurich

- **Topic ID:** 01
- **IVDA Research Stream:** Item Ranking, Interactive Machine Learning
- **Domain:** Short-Term apartment renting (Airbnb). The anticipated scenario for this group project is about users who want to find interesting apartments for rent in Zurich, using Airbnb as an example.
- **Dataset:** zurich apartment renting.csv
- **Dataset Type:** Multivariate (num, cat, mixed), 2252 items (apartments), 18 attributes.
- **User group:** Potential Airbnb customers in Zurich. Non-experts, with individual preferences.
- **Motivation/Goal:** Goal is to enable individual users to find the apartment listings that match their preferences best. The underlying interactive ML principle will enable users to give scores to individual Airbnb listings, and a regression model will, after obtaining sufficient training data, output predictions that are aligned with expressed user preferences, submitted through labels. An explainable AI extension would include displaying the weights/contributions of each attribute/feature to the regression computation, to understand the model behavior. The interactive item labeling approach will suffer from cold start problems: at the beginning, no labels are given, i.e., the regression model cannot yet produce a (meaningful) output (rating predictions) for unlabeled apartments. Several principles can be included to overcome this situation:
  - Calibration: inquire the user to submit a minimum set of score labels to a small subset of items/instances.
  - Active Learning: the suggestion of instances (by the system) to be labeled/scored by users, aiming at improving the performance of the regression model by meaningful instance selections.
  - Gamification: to motivate users to submit yet more labels (training data). This would mitigate the problem that human labor is typically tedious and boring. Game design elements (points, badges, performance graphs, competition, etc.) need to be coupled with active learning and quality improvement metrics to assess and improve the learning progress (training data size, confidence of the learner, reduction of error, etc.). Telling the user how certain the model is could also be a gamification idea: a gamified increase of model certainty.

  Note that, similar to most personalized learning scenarios, no ground truth data exists that can be leveraged. To limit the scope, it is not expected that you implement all three principles.

Interactive Exploration of a Catalog of Digital Editions

- **Topic ID:** 02
- **IVDA Research Stream:** Data Exploration, Exploratory Search
- **Domain:** Digital Library
- **Dataset:** See Link
- **Dataset Type:** multivariate, about 40 mixed data attributes.
- **User group:** Researchers, Digital Library Users, Historians
- **Motivation/Goal:** The *Catalog of Digital Editions* is a metadata collection of 350 works from Digital Humanities projects, manually curated by Greta Franzini. The solution will be in the realm of exploratory search: combining open-ended information-seeking behavior (exploration) with directed search and exploration support. Think about how interactive visualizations can allow users to explore the data. Choose appropriate visualizations for the different facets (e.g., time, geo, categorical, numerical, binary) and combine them in an interactive interface. Make appropriate assumptions in cases where the features in the dataset are not self-explanatory. If interested, you can also think of LLM-Based Data Enrichment for yet other, non-existing attributes/metadata.
- **Note**: Please refer to our announcement about the LLM-Based Data Enrichment Special, above the → Project Topic Descriptions.

Sector and Industry Group Similarity of Companies

- **Topic ID:** 03

- **IVDA Research Stream:** Exploratory Data Analysis, Relation Discovery, Hierarchical Data Analysis,
- **Domain:** Finance
- **Dataset:** Stock Sectors Dataset (LLM-Generated)
- **Icons per Stock:** Icons, Logos
- **Dataset Type:** Mixed. ID can be ignored; "confidence" means the confidence of the LLM in the dataset generation. Use 'ISIN' as the primary key for companies. Sectors, Industry Groups, and Industries form a hierarchy (important for data exploration).
- **User group:** Finance experts
- **Motivation/Goal:** An analyst received sector information from an API, where each company is assigned to one of 11 sectors according to the Global Industry Classification Standard (GICS). However, the 11-sector classification is binary and too coarse. For example, NVIDIA (a CPU/semiconductor company) and Uber are both labeled as "Information Technology", even though they clearly do not belong to the same peer group. To obtain more fine-grained sector information, the expert applied an LLM-based scoring approach to all stocks, spanning two hierarchical levels: sectors (11 categories) and industry groups (25 categories). Each company now distributes across multiple relevant categories at each level, summing to 100%. For instance, NVIDIA's sector information is represented as 70.8% Information Technology, 20% Consumer Discretionary, and 9.2% Communication Services. Beyond the binary classification, the analyst can now explore stock data as a continuum, opening up new exploration opportunities. But how? You are asked to provide an IVDA solution. The analyst needs a nearest-neighbor analysis (similarity metric) and stock exploration in the context of the sector hierarchies. Also, the analyst does not trust the LLM-generated data blindly: some insight into the quality of the data (on demand) would be desirable. Finally, the analyst has disclosed information about owned stocks, opening up opportunities to assess the diversification of the portfolio with respect to sector information. Does the analyst need to worry about a biased portfolio towards some sectors? To keep the scope of the project limited, the minimum target is to include sectors (11 categories) only. Additionally, including the industry groups (25 categories) as the second hierarchical layer is optional; if included, the analyst reminds you of the hierarchical structure of the sector and industry group data.

Item Ranking for Stocks

- **Topic ID:** 04
- **IVDA Research Stream:** Interactive Item Ranking, Data and Preference-Based Decision-Making
- **Domain:** Finance. Professional investors in financial markets, as well as interested non-expert stakeholders. From the many possible types of market analyses, we focus on an item ranking scenario, i.e., the interactive creation of stock rankings, by meaningful multidimensional criteria. Rankings of stocks are used in a variety of cases, including portfolio diversification, investment strategy, performance evaluation, or risk management. In essence and in contrast to gut-feeling decisions, we want to study a systematic approach to stock ranking, and thus will design and develop an IVDA approach.
- **Dataset:** f0f1f2Data.csv
- **Icons per Stock:** Icons, Logos We limit the scope to 1000 relevant stocks, i.e., 1000 of the largest companies listed on stock exchanges all over the world.
- **Dataset Type:** multivariate (num, cat, mixed)
- **User group:** Stock market enthusiasts, investors, experts, but also and non-experts.
- **Motivation/Goal:** In contrast to expensive and extensive power-user and expert tools such as Bloomberg terminals, this group project will focus on the design and development of an IVDA tool that is simple enough to be used by non-experts. Users will be able to interactively rank 1000 stocks, by 13 temporal attributes serving as ranking criteria. The novelty of the approach lies in an innovative simplification: Users will be enabled to rank stocks by three criteria for every temporal attribute:
  - f0: the current value of an attribute (such as the current *eps*, *returnOnAssets*, or *debtEquityRatio*)
  - f1: the first-order derivative/growth (such as the growth of the revenue over time – this is an indication that a price may also change)

  - f2: the second-order derivative/momentum/change-of-change (such as the acceleration of the revenue value change, possibly indicating early changes)

The goal of this project is to enable users to rank stocks interactively, by these 3×13 criteria, in arbitrary combinations. The overall ranking shall be computed by ordering stocks by their weighted sums of the individual attribute rankings. Weights to be considered can be per attribute, per f0/f1/f2 priority, and/or both. The interactive ranking-creation component forms an integral part of an overview-to-detail approach to stock analysis. A final challenge regards the assessment of change: did a stock rise/fall in the ranking due to adding/changing ranking preferences by the user?

Personalized Ranking for Universities (incl. LLM-Based Data Enrichment)

- **Topic ID:** 05
- **IVDA Research Stream:** Interactive Item Ranking, Data and Preference-Based Decision-Making, LLM-Based Data Enrichment
- **Domain:** Research and Education. University rankings aim to sort/order universities. The advantage is compatibility and decision-making, often including universities worldwide. The disadvantage of current university rankings is their strong focus on basic KPIs. Some people say, these rankings are rather mechanical and tend to ignore more societal and human values. Also, most third-party rankings are pre-computed and cannot be adjusted with respect to user preferences, such as the weights of characteristics/attributes.
- **Dataset:** University Ranking The universities are distributed all over the world. Based on this standard university ranking dataset, part of the process is that you ideate five more attributes and gain values through IVDA's LLM-based data enrichment method. You will meet Yves, who will guide you through this process.
- **Dataset Type:** multivariate (num, cat, mixed)
- **User group:** Stakeholders involved in the research and teaching ecosystem: students, professors, researchers, or university decision-makers.
- **Motivation/Goal:** In contrast to the rather static and mechanical, KPI-driven university rankings, this project seeks to ideate five characteristics/attributes about the World's universities that inherently focus on human values for the good. While it is up to the student group to make a final decision on these five attributes, they should comply with the targeted stakeholder group. Example: a university decision-maker may be interested in the "Sustainability Development Impact", or similar. The goal of the project would be to a) present the existing ranking, introduce five new attributes that are clearly visible as such in the tool, and enable control for stakeholders to create an overall university ranking. Stakeholders may also be interested in validating the plausibility and trustworthiness of the five LLM-generated attributes - this can be supported by a view that reveals attribute relations for plausibility checks.
- **Note:** Please refer to our announcement about the LLM-Based Data Enrichment Special, above the → Project Topic Descriptions.

Country Sustainability Development Analysis (incl. LLM-Based Data Enrichment)

- **Topic ID:** 06
- **IVDA Research Stream:** Exploratory Data Analysis, Interactive Relation Discovery, LLM-Based Data Enrichment
- **Domain:** Sustainability Research
- **Dataset:** Country Sustainability and Forest Datasets
- **Dataset Type:** temporal, multivariate (num, cat, mixed). Starting off from a basic dataset containing the time series data of 20 years for a series of countries and all 17 SDG goals, this project will study two analytics aspects. First: exploring the temporal development of SDG scores per country (and as an extension: aggregated groups of countries), and second: discovering relations between current SDG scores (as of 2022) and external attributes/features/dimensions that help to contextualize and explain SDG scores of a country. An example for the latter could be: with the IVDA tool, users can clearly see that high SDG scores for SDG Goal X relate to high forest coverage in these countries. To enable such contextualizations and explanations of SDG scores, the student team is asked to integrate external attributes/features/dimensions, using countries as a primary key to join the basic dataset with external attributes. These attributes may be found in external datasets on the web or be ideated

through LLM-Based Data Enrichment. An example prompt for the latter would be: ”give me, for countries X, Y, Z, ..., the average temperature value in January, as well as the model confidence, and a short textual explanation for the retrieved values”. Examples of external attributes (not limited to) are GDP, forest coverage, type of government, type of dominating religion, continent, etc. For the first part of the IVDA approach, all these external attributes can be used for grouping the SDG score time series per country, e.g., all aggregated countries for continents, in a small multiples arrangement.

- **Note:** Please refer to our announcement about the LLM-Based Data Enrichment Special, above the → Project Topic Descriptions.

Fatigue Data - Interactive User-Centered Approach

- **Topic ID:** 07
- **IVDA Research Stream:** Personalized Data Analytics, Time Series Prediction, Recommendation
- **Domain:** sensor data analysis, fatigue management, personal data analytics
- **Dataset:** Dataset Link (Zenodo)
- **Dataset Type:** timeseries, text
- **User group:** end users (looking to manage their fatigue)
- **Motivation/Goal:** This combination of continuous multimodal wearable sensor data and the daily fatigue questionnaires offers a unique opportunity to predict and understand fluctuations in fatigue levels over time. Leveraging AI, we could harness this data to develop a proactive system that provides a real-time fatigue assessment from the end-user, enabling timely interventions and personalized recommendations. Our motivation stems from the potential to empower individuals to actively manage their fatigue and enhance their quality of life. You could also imagine that such a tool may provide valuable insights to users' healthcare teams, allowing for more effective patient care. The dataset includes 28 subjects, over 973 days of sensor and survey data collection. You can read more about the study that collected this data here: Assessment of Fatigue Using Wearable Sensors: A Pilot Study

  The goal here is to create an AI-driven system that combines multimodal wearable sensor data with daily fatigue patient-reported outcomes (PROs). The system should employ ML techniques to predict and visualize fluctuations in fatigue levels, allowing for timely interventions and personalized recommendations to improve individuals' well-being. This problem statement could also be adapted to a human-model teaming approach, if desired.

Visual Analytics for Activity-Aware Diabetes Self-Management

- **Topic ID:** 08
- **IVDA Research Stream:** Personalized Analytics, Disease Self-Management, LLM-Based Data Enrichment, Relation Discovery, Pattern and Trend Analysis
- **Domain:** Healthcare, Type 1 Diabetes.
- **Dataset:** BrisT1D-Open Dataset (subset, if possible request the full dataset)
- **Dataset Type:** Time-series (continuous glucose, insulin, meals, heart rate, steps, calories, activity labels) + textual (interview/focus group transcripts)
- **User group:** Individuals with Type 1 Diabetes (T1D)
- **Motivation/Goal:** The BrisT1D dataset combines continuous multimodal wearable sensor data (heart rate, steps, calories, distance) with detailed records of insulin delivery, carbohydrate intake, and continuous glucose monitoring from young adults with Type 1 Diabetes. Alongside these physiological signals, monthly interviews and focus group transcripts provide qualitative insights into how participants manage their condition and use smartwatches in daily life.

  This combination of continuous physiological data and personal narratives is quite unique, allowing users to understand how physical activity influences blood glucose regulation and self-management behaviors. By applying standard NLP methods and/or LLM-Based Data Enrichment to the transcripts - such as topic and sentiment extraction - the unstructured qualitative data can be transformed into analyzable features that complement the quantitative streams. Together, these data sources can reveal individualized patterns of glucose response to activity and uncover how people perceive and adapt their management strategies.

The goal of this project is to create an interactive visual analytics solution that integrates wearable data with LLM-enriched qualitative insights to help users explore and interpret how activity affects their glucose control. The system should visualize relationships between exercise, heart rate, and blood glucose, and allow users to examine patterns such as post-exercise glucose drops or activity-specific differences. The customizable IVDA solution will enable users to focus on what matters most to them (e.g., understanding daily routines and their perceptions of them or comparing multiple activity types). ML models (e.g., clustering or regression) can be used to identify recurring activity–glucose patterns and predict glucose trends. Please also find a description of the dataset here.

- **Note**: Please refer to our announcement about the LLM-Based Data Enrichment Special, above the → Project Topic Descriptions, but you can also use NLP methods within your framework to facilitate AI-based data enrichment.

Landscape Decision Spaces

- **Topic ID:** 09
- **IVDA Research Stream:** Data and Preference-Based Decision-Making, LLM-Based Data Enrichment
- **Domain:** Environment
- **Dataset:** Ask Martin
- **Dataset type:** multivariate (geo, seq, num)
- **User group:** Stakeholders, Policymakers, Experts
- **Motivation/Goal:** Ecosystem services (ES) are the provisioning (e.g., food), regulating (e.g., climate), cultural (e.g., recreation), and supporting (e.g., nutrient cycling) benefits that humans derive from nature. Understanding the tradeoffs between landscape configurations (e.g., development plans, agricultural buffers) and corresponding ES may help to align actions with desired outcomes. Create a visual analytics tool that helps users avoid or seek specific changes in ES to serve as discussion points and enhance communication (https://doi.org/n49c) between and among stakeholders, policymakers, and experts.

Medical Diagnostics

- **Topic ID:** 10
- **IVDA Research Stream:** Doctor-Patient Interaction, LLM-Based Data Enrichment, Explainable AI
- **Domain:** Healthcare
- **Dataset:** Kaggle Simulated Medical Interviews
- **Dataset type:** textual (recorded talks of doctor and patients), which should be transformed into structured data, meaning a tabular dataset with symptoms, diagnosis, chain of thought from symptoms to diagnosis and suggested treatments (all cactegorical). To get to structured data, a possibility could be to use the LLM-Based Data Enrichment Special.
- **User group:** Physicians
- **Motivation/Goal:** Physicians spend a lot of time documenting their interactions with patients. Obviously, this should not be the main job of physicians. What would help physicians a lot would be a tool to support them in the documentation and decision-making process after patient talks. Therefore, the goal of the IVDA tool would be to assist physicians by extracting and organizing key information from doctor-patient dialogues. For example, integration or preprocessing through an LLM could help extract and/or highlight key information (symptoms, diagnosis, suggested, chain of thought from symptoms to diagnosis, treatment). An important factor for physicians is transparency. Therefore, the IVDA tool should support exploring the symptoms and the chain of thought, so they have a clear picture how the LLM got from symptoms to diagnosis and suggested treatment. In addition, doctors would also be able to edit or override the output (suggested diagnosis, treatment, etc.)

  Note: Please refer to our announcement about the LLM-Based Data Enrichment Special, above the → Project Topic Descriptions, but you can also use NLP methods within your framework to facilitate AI-based data enrichment.

Personalized Job Ranking for Career Path Exploration

- **Topic ID:** 11
- **IVDA Research Stream:** Interactive Item Ranking, Multi-Criteria Decision Making
- **Domain:** Career Development
- **Dataset:** LinkedIn Job Postings Dataset (2023-2024)
- **Dataset type:** tabular, multivariate (num, cat, text, mixed). The student group is encouraged to sample jobs down to 1000 (try to avoid biases).
- **User group:** Students, career counselors advising, career switchers
- **Motivation/Goal:** The job search process is overwhelming for students and early-career professionals who must evaluate hundreds of opportunities across multiple dimensions: compensation, location, required skills, growth potential, work-life balance, and company culture. Traditional job sites just show endless lists sorted by date, forcing job seekers to mentally juggle dozens of factors at once. Many people don't even know what they really want until they see it. This IVDA tool will help users explore jobs interactively, compare what matters most to them, and discover opportunities they might have missed, all while learning what they value in a career. In case the group finds out that an important attribute is missing in the dataset, it is possible to leverage the IVDA Group's LLM-Based Data Enrichment Method to add this attribute to the dataset.
- **Note**: Please refer to our announcement about the LLM-Based Data Enrichment Special, above the → Project Topic Descriptions.

Healthy Recipe Ranking and Personalized Nutrition Menu Planner

- **Topic ID:** 12
- **IVDA Research Stream:** Interactive Item Ranking, Multi-Criteria Optimization, Personalized Recommendation
- **Domain:** Dietary Management
- **Dataset:** Epicurious Recipes with Rating and Nutrition
- **Dataset type:** tabular, multivariate
- **User group:** Health-conscious millennials planning weekly meals
- **Motivation/Goal:** Choosing what to eat shouldn't feel like solving a math problem, but it often does. People want meals that are healthy, delicious, quick to make, affordable, and fit their dietary needs all at once. A diabetic athlete needs high protein and low carbs. A busy parent needs quick meals that the whole family will eat. Current recipe apps force you to pick one priority: ”healthy” OR ”fast” OR ”budget-friendly.” This tool lets users balance all their needs at once, see the trade-offs clearly (yes, the healthiest option takes 20 minutes longer), and plan a week of meals that actually work for their real life, not some idealized version of it. In case the group finds out that an important attribute is missing in the dataset, it is possible to leverage the IVDA Group's LLM-Based Data Enrichment Method to add this attribute to the dataset.
- **Note**: Please refer to our announcement about the LLM-Based Data Enrichment Special, above the → Project Topic Descriptions.

UN General Debate: Visualizing Rhetorical Priorities (incl. LLM-Based Data Enrichment)

- **Topic ID:** 13
- **IVDA Research Stream:** Temporal and Comparative Text Visualization, LLM-Based Data Enrichment
- **Domain:** International relations and political communication. UN General Debate speeches reflect shifting national priorities and global narratives across decades.
- **Dataset:** UN General Debate Corpus (1970–2016)
- **Dataset Type:** textual (speeches) + categorical (country, year, region). The student group is encouraged to sample the data objects down to 1000 (try to avoid sampling biases).
- **User group:** Political scientists, diplomats, journalists, students of international affairs

- **Motivation/Goal:** Present a baseline exploration of speeches and *introduce five new human-centric attributes* (e.g., *security urgency*, *cooperation vs. confrontation*, *development emphasis*, *human-rights framing*, *multilateralism tone*). Build a visual analytics tool to explore these attributes across countries/regions and over time (filtering, comparisons, small multiples, heatmaps), and include a plausibility view that relates enriched attributes to representative passages.
- **Note**: Please refer to our announcement about the LLM-Based Data Enrichment Special, above the → Project Topic Descriptions.

Craigslist Cars: Visualizing Lifestyle Fit (incl. LLM-Based Data Enrichment)

- **Topic ID:** 14
- **IVDA Research Stream:** Attribute Exploration and Trade-off Visualization, LLM-Based Data Enrichment
- **Domain:** Consumer transportation and marketplace analysis. Beyond specs, buyers care about lifestyle fit and everyday usability.
- **Dataset:** Craigslist Cars and Trucks
- **Dataset Type:** mixed (numeric, categorical, listing text). The student group is encouraged to sample data objects down to 1000 (try to avoid sampling biases).
- **User group:** Car buyers, mobility researchers, marketplace stakeholders
- **Motivation/Goal:** Present a baseline view of key specs and *introduce five new lifestyle attributes* (e.g., *urban practicality*, *long-trip comfort*, *family friendliness*, *eco-image*, *maintenance risk*). Build a visual analytics tool that lets users explore trade-offs between enriched attributes and the raw specifications about cars provided in the dataset. Enable information drill-down through faceted filtering, and data exploration through multiple linked views, possibly including scatter plots or parallel coordinates. The result of a clustering routine may reveal groups of cars, while dimensionality reduction/embedding in 2D may provide a similarity-preserving perspective to cars. For the LLM-generated attributes, the question arises as to how plausible this data is. The visualization of model confidences for generated values may be a line of approach here.
- **Note**: Please refer to our announcement about the LLM-Based Data Enrichment Special, above the → Project Topic Descriptions.

Exploring the UCI Heart Disease Dataset

- **Topic ID:** 15
- **IVDA Research Stream:** Health Data Exploration, Relation-Discovery, Unsupervised Machine Learning
- **Domain:** Healthcare
- **Dataset:** UCI Heart Disease Dataset
- **Dataset Type:** Mixed, well-known dataset consisting of medical records of patients with multiple attributes such as age, sex, chest pain type, blood pressure, cholesterol level, and binary heart disease diagnosis.
- **User group:** Healthcare professionals
- **Motivation/Goal:** The goal is to create an IVDA tool that enables users to explore and investigate relationships in the data. You will design an IVDA tool that supports traversal between items and attributes. An item refers to an individual patient record, while attributes refer to the patient characteristics (such as age, blood pressure, and cholesterol level).

  A working principle of this project is the idea that users can express item-based questions and receive attribute-based responses, and vice versa. Along these lines, the exploratory tool should enable users to identify an item or an item subset of interest, and receive answers to interesting questions in the attribute space, such as: "Are this focused patient's characteristics typical of others with heart disease?" and provide ways to investigate corresponding items that help answer such questions interactively. Additionally, the IVDA tool should enable attribute-to-item traversal, in which users can specify attribute values of interest and reveal relevant patients, for downstream analysis and comparison, helping to answer the attribute-based question. The broad goals of this project are to learn to design linked views that support navigation between different facets of data, understand techniques to implement item-to-attribute and attribute-to-item traversal, and incorporate both data-driven and human-driven discovery in visual analytics support. Feel free to incorporate additional datasets, add filtering, clustering, or dimensionality reduction views, and integrate LLMs.