Group Assignment 1 (G1): Project Characterizations

Interactive-Visual Data Analysis
Fall 2023

DUE: November 5, 23:59

General Task Description

You have now transitioned to the group project phase of the IVDA course. Here, you will form groups of 4 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. You will be guided, through the group assignments and future exercises, in how to build a VA tool from scratch. The work you have already done in A1-3 will be quite helpful in this effort, as well. The project will done with a group you will form together now, and stick with until the end of the term. Our goal is to give each group the opportunity 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 get to develop a powerful skill-set in IVDA development.

Your first task is to identify a problem of interest to your group from our provided list of possible problem statements, and to characterize your problem space using the **what-why-how** method presented in class. This assignment assumes that you have already declared the 3- or 4-person group of students you will work with to the course team.

Your group form must be filled by **Wednesday, October 25th, 23:59**: `https://forms.gle/XtdL9v4y6CXPWzrK7`

Point Distribution

This assignment is worth 10 points in total, for each student. All answers should be related to your group's selected problem statement. Generally, the points are allocated as follows, with a more detailed breakdown in the Written Response Instructions:

- **What:** Characterization of your project (3 points)
- **Why:** Characterization of your project's tasks (3 points)
- **How:** Proposed methodologies for supporting the tasks identified in the previous step (3.5 points)
- **Group Dynamics:** Describe the roles each of your group members will fulfill (0.5 points)
- **References and in-text citation (in written response)** (mandatory)

Submission

Your written response should include:

- 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 4 of you). Your OLAT directory will become available to you once we have registered your group and problem statement choice (by October 25, 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 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. Assignments without proper references (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** (appx. 1 page). Please use the provided template to prepare your submission, and include any references that you used:

What

1. Describe your project's domain and problem. You should be referring to, **and then expanding on,** the information in your topic's problem statement (0.5 points)
2. Characterize the attributes of the data source you intend to use in your project, according to Munzner's VAD Chapter 2 (0.5 points)
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 you expect to do in future phases of your project, especially in how they relate to your answers in the **How** section of your response below (2 points)

Why

1. Describe your project's target user(s). You should be referring to, **and then expanding on,** the information in your topic's problem statement (0.5 points)
2. Characterize the core analysis tasks you believe your project is meant to support, according to Munzner's VAD Chapter 3. Typically, the projects should not include more than 5 tasks (1 points)
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 (1.5 points)

How

1. Describe your proposed visualization for supporting the 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). You may wish to use the **(What + Why) = How** sentence structure discussed in the lecture and exercise sessions (2 points)
2. Include a sketch, done either by hand or using Figma. Consider the view composition and interactions that would assist in supporting your identified tasks (L05). Make clear annotations to describe the interactions you have in mind (1 point)
3. Describe (briefly!) your proposed modeling 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) (0.5 points)

Group Dynamics

Using the Data Baton definitions, 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)

Problem Statements - 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 first-come, first-served, so please pick out at least 3 choices your group would be happy to work on. We will distribute your topics by October 26th, 2023, 14:00.

The problem statement you are assigned will also be the topic which you'll characterize in your G1 submission. The possible IVDA topics for the group project are in the list included below. Some of these topics may have more than one possible problem statement to choose from:

1. XAI for Recommender Systems
2. IVDA for Data Humanism
3. Human-AI Teaming
4. Human-Centered Healthcare
5. Personalized Healthcare
6. Item Ranking
7. Item Labeling
8. Digital Libraries
9. Sustainability
10. Item Similarity, Search and Exploration
11. Human-Centered Relation Discovery

XAI for Recommender Systems

- **Topic ID:** 01
- **Domain:** social media (dating)
- **Dataset:** See Sharepoint
- **Dataset Type:** tabular, text, catgorical/ordinal, categorical/nominal
- **User group:** model developer
- **Motivation/Goal:** In the era of digital connections, understanding how data from various users can be effectively combined is crucial in enhancing the accuracy and quality of recommendations. Reciprocal recommender systems offer a unique and dynamic environment for studying the fusion of user data, as the choice of fusion method significantly impacts the recommendations made to users. How can data visualization be used to explain different fusion approaches in the context of a reciprocal recommender system? (better for those already experienced with RecSys or highly skilled in ML)
- **Topic ID:** 02
- **Domain:** gaming
- **Dataset:** See Sharepoint
- **Dataset Type:** tabular
- **User group:** end user
- **Motivation/Goal:** As the Pokémon franchise continues to evolve and expand, the importance of making informed decisions when selecting a Pokémon for one's team has become increasingly paramount. With an ever-growing roster of Pokémon species and a multitude of personal preferences, gamers are faced with the challenge of assembling the most effective and enjoyable team. To address this need, there arises a demand for innovative solutions that can aid Pokémon enthusiasts in optimizing their team compositions. How a visual analytic tool can empower gamers to make well-informed decisions about which Pokémon to include in their team? (you can use the Turi-create package from Apple)
- **Topic ID:** 03

- **Domain:** sport
- **Dataset:** See Sharepoint
- **Dataset Type:** tabular
- **User group:** coaches
- **Motivation/Goal:** As the world of FIFA continues to evolve and diversify with the introduction of both male and female players, the importance of informed decision-making in forming a mixed-gender team has never been greater. With an extensive pool of talented players across the top 5 European leagues (Italy, Germany, Spain, France, and the UK), FIFA enthusiasts are faced with the exciting yet challenging task of assembling a competitive and enjoyable mixed-gender team. To address this need, there arises a demand for innovative solutions that can empower FIFA enthusiasts to optimize their team compositions. The goal of this project is to develop a visual analytic tool that leverages the FIFA male and female player datasets from the top 5 European leagues (Italy, Germany, Spain, France, and the UK) to help FIFA enthusiasts create their first competitive and diverse mixed-gender team. The tool should provide insights and recommendations for player selection, considering factors such as player attributes, performance statistics, and compatibility within the team. By doing so, this project aims to enhance the FIFA gaming experience, promote inclusivity, and encourage players to explore the diverse talent pool available in the FIFA video game.

IVDA for Data Humanism

- **Topic ID:** 04
- **Domain:** social media (dating)
- **Dataset:** See Sharepoint
- **Dataset Type:** tabular + text
- **User group:** end user
- **Motivation/Goal:** Textual explanations provided for recommendations are often not interactive and not personalized for the user. This shortcoming is a critical concern, considering that data is not merely a collection of static artifacts but rather a representation of human behavior. The concept of data humanism strives to transform data visualization into a personalized experience, emphasizing that data is intricately linked to the essence of human existence. At its core, data humanism advocates for nonlinear storytelling within data visualization, allowing individuals to immerse themselves in the details, lesser-known narratives, and overarching trends that data can unveil. How can data visualization techniques, especially not classical ones, together with a storytelling approach explain a recommendation algorithm? (TIPS: for the recommended system use a content-based strategy, you could leverage the **Turi-Create** package from Apple)

Human-AI Teaming

- **Topic ID:** 05
- **Domain:** athletics
- **Dataset:** `https://www.kaggle.com/datasets/niharika41298/gym-exercise-data`
- **Dataset Type:** tabular, text, categorical/ordinal, categorical/nominal
- **User group:** self-training athletes, coaches, injured athletes, or physiotherapists
- **Motivation/Goal:** This dataset provides a list of gym exercises and their target muscle groups. This information could be really valuable for individuals who train regularly in some sport that requires the coordination of multiple muscle groups, and requires each of those to be strong (think running, climbing, swimming, tennis, football, rugby, golf, archery...). It could also be valuable for someone just getting into a new sport and hoping to avoid injury, or someone who is already injured and looking to strengthen their surrounding muscle groups. You could also consider incorporating factors like injury history, training intensity, or game schedules. Your goal here would be to develop a VA application that uses an interactive ML approach to assesses an athlete's skill level, provide targeted skill-building exercises, and track their progress. The objective is to optimize skill development and boost performance in a way that takes into account the insights of both the user, and the model.

Human-Centered Healthcare

Fatigue Data - Biomarker Identification

- **Topic ID:** 06**a**
- **Domain:** sensor data analysis, fatigue management, clinical trial analytics
- **Dataset:** `https://zenodo.org/records/4266157`
- **Dataset Type:** timeseries, text
- **User group:** researchers, healthcare professionals
- **Motivation/Goal:** In an era of comprehensive health monitoring through wearable sensors, having this multimodal data, combined with corresponding daily fatigue questionnaires holds great potential for shedding light on the factors influencing daily fatigue levels. However, the complexity of this data source requires a specialized visual analytics platform. Our motivation is to provide researchers and healthcare professionals with a powerful tool to explore this multimodal information, enabling them to uncover temporal patterns and potential fatigue risk factors, thus improving the management of fatigue-related issues. 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 develop a VA platform that integrates multimodal wearable sensor data and the daily fatigue questionnaires into a platform that enables researchers to visualize and analyze temporal patterns, relationships, and risk factors associated with fatigue, facilitating the identification of key factors influencing daily fatigue levels.

Fatigue Data - Interactive User-Centered Approach

- **Topic ID:** 06**b**
- **Domain:** sensor data analysis, fatigue management, personal data analytics
- **Dataset:** `https://zenodo.org/records/4266157`
- **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 providing 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 machine learning 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.

Personalized Healthcare

- **Topic ID:** 07
- **Domain:** healthcare
- **Dataset:** provided after NDA
- **Dataset Type:** time series
- **User group:** end user (individuals with T1D)
- **Motivation/Goal:** The recent advances in self-monitoring technologies empower individuals to benefit from the abundance of personal health data. This is particularly important for chronic disease self-management, where effective self-management is crucial to avoid severe health complications or even death. Designing technologies for chronic disease self-management is challenging as their end users are often heterogeneous in terms of visual understanding, age, lifestyle, etc. However, current patient-facing technologies for chronic disease management are typically designed using a one-size-fits-all approach, generalizing the nuances of users' very complex and individual disease management practices.

  One condition that would particularly benefit from a more individualized technological approach is type 1 diabetes (T1D). T1D is a chronic autoimmune disease that relies on lifelong exogenous insulin supplementation to regulate blood glucose values. T1D typically occurs in childhood or adolescence and affects about 9 million people worldwide. Treatment of T1D is particularly challenging due to the multitude of factors affecting blood glucose control such as insulin administration, food consumption and absorption, physical activity, stress, and hormonal levels. Self-management in T1D is particularly personal as the body reacts individually different to factors affecting blood glucose levels and the end users of technology are heterogeneous regarding factors such as lifestyle, age, preferences of management involvement or even technologies used for management.

  The goal of this project is to develop a visual analytics tool for individuals with T1D that individually adapts at least one aspect. Whether you want to focus more on personalization addressing differences regarding visual literacy, user preference (e.g., of interaction/visualization), age, lifestyle, or something you deem interesting is up to you.

Similarity for Multidimensional Time Series

- **Topic ID:** 08
- **Domain:** healthcare
- **Dataset:** See Sharepoint
- **Dataset Type:** time series
- **User group:** end user (individuals with T1D)
- **Motivation/Goal:** Informed decision-making based on an assessment of one's current situation and experience is essential in chronic disease self-management, like type 1 diabetes (T1D) self-management. The importance of effective decision-making lies in its effect on a patient's quality of life and health outcome. T1D self-management decision-making is particularly difficult as effective management depends on a multitude of interdependent factors including current blood glucose level, active insulin, active carbohydrates, and stress. While most individuals with T1D rely on their past experiences to determine what to do in their current situation, human memory is usually limited and often biased. This may then lead to misinformed decision-making and worse health outcomes. Thus, the goal of this project is to implement a visual analytics tool that allows users to 1) select a "current" situation of interest 2) use the multidimensional diabetes data to identify the most similar past situation and 3) visualize those situations to the user.

Item Ranking: S&P 500 Stocks

- **Topic ID:** 09
- **Domain:** Finance. Not only our faculty is interested in finance markets, but also many stakeholders are. From the many possible types of market analyses, we focus on an item ranking scenario, i.e., the interactive creation of stocks orderings by meaningful multidimensional criteria. Ranking stocks is 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 identification through interactive stock ranking.

- **Dataset:** See Sharepoint We limit the scope to 500 S&P stocks, i.e., 500 of the largest companies listed on stock exchanges in the United States.
- **Dataset Type:** temporal, 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 the S&P 500 stocks, by multiple temporal attributes serving as ranking criteria. The novelty of the approach lies in a simple assumption: Users will be enabled to rank stocks by three criteria for every temporal attribute:
  - f0: the basic values of attributes (such as the market capitalization, e.g., the higher, the better)
  - f1: the first-order derivative/growth (such as the growth of the revenue over time)
  - f2: the second-order derivative/momentum (such as the acceleration of research and development investments of a company in the very moment)

  Goal of this project is to enable users to rank stocks interactively, by these $3 \times attributeCount$ criteria, in arbitrary combinations. The overall ranking shall be computed by ordering stocks by their weighted sums of the individual attribute rankings. It is intended that this ranking component forms an integral part of an overview-to-detail approach for stock analysis.

Item Labeling: Apartment Ratings in Zurich

- **Topic ID:** 10
- **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:** See Sharepoint
- **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, a regression model will, after obtaining sufficient training data, output predictions that are aligned with expressed user preferences, submitted through labels. 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. Two principles can be included to overcome this situation:
  - Active Learning: the meaningful selection of instances to be labeled next, aiming at improving the performance of the regression model. Strategies for instance selection may be based on model criteria (instances of high uncertainty), data criteria (regions of the dataset that have not yet been labeled), or user-based criteria (the user has identified a pattern in the data that deserves to be labeled next)
  - Gamification: to motivate the user to submit more labels (create training data), mitigating the problem that human labor is typically tedious and boring. Game design elements (points, badges, performance graphs, competition, etc.) need to be coupled with metrics to assess the learning progress (training data size, confidence of the learner, reduction of error, etc.).

  Note that, similar to most personalized learning scenarios, no ground truth data exists that can be leveraged.

Digital Libraries: Temporal Locations of Persons in Bullinger's Network in the 16th Century

- **Topic ID:** 11
- **Domain:** Digital Libraries. Digital editions are recreations of historical documents/artifacts that have been carefully curated and edited using library methods such as metadata indexing and formalization through markup standards such as TEI-XML. Digital editions can vary from the very minimal to the very complex, allowing for many forms of

interactions, often concerning named entities, places, images of the originals, diplomatic or reading transcriptions, historical variations, and other types of unifying characteristics.

The digital editions provided by the Zentrum Digitale Editionen & Editionsanalytik in Zurich are a core application of the Digital Humanities. Disciplines such as history, philology, and law make digitally edited sources available online to make them searchable and usable for further research questions. A digital edition that is particularly popular in the Zurich area is the Bullinger Edition. Heinrich Bullinger was a prominent Swiss writer and reformer, his letters offer unique insights into his life and travels, situated in the 16th century. More than 10,000 of his letters have already gone through a systematic digitalization and transcription process.

- **Dataset:** in preparation. **If interested, please speak to course team in addition to filling out the group declaration form!**
- **Dataset Type:** tabular, temporal, geo, multivariate, (num, cat, mixed). Temporal locations of 50+ persons of Bullinger's network, retrieved from more than 10,000 original letters.
- **User group:** Historians, linguists, researchers, general public library users
- **Motivation/Goal:** With the digitalization of Bullinger's letters, great opportunities emerge for large user groups to engage with this cultural heritage: people may gain access to the documents, apply novel forms of search and exploration activities, identify documents of particular interest in huge corpora to be re-used for the in-depth study, e.g., for scientific purposes. IVDA can play a central role in supporting these user groups with new forms of interactive engagement with these digital editions.

  We are interested in leveraging the geographical and temporal information provided by Bullinger's letters. Given the number of letters and the different persons mentioned, it is still challenging to understand the complex connections and relations of these individuals in the 16th century, exposed by their temporal locations, and peoples' mobility. An IVDA tool will allow the exploration and detailed analysis of historical travel information of Bullinger's network. This will reveal valuable information and relations, which would be tedious to achieve through text analysis and traditional manual document screening. The IVDA tool is expected to provide overview perspectives, browsing and drill-down interactions, as well as a detailed analysis of identified patterns.

Sustainability

UZH Publications and SDGs

- **Topic ID:** 12
- **Domain:** Academic Research and Sustainability
- **Dataset:** See Sharepoint
- **Dataset Type:** 27162 rows $\times$ 202 attributes; multivariate (num, cat, mixed)
- **User group:** Academic environments and societies, Educational institutions, Students, Professors, NGOs
- **Motivation/Goal:** This problem statement is motivated by our ongoing research project, the SDG Research Scout. This project is dedicated to harnessing the power of cutting-edge NLP, AI, and interactive visual data analysis to identify content in scientific publications related to the UN Sustainable Development Goals. By developing a system that can decipher the details of SDG definitions at both goal and target levels, we aim to make it easier for UZH members to engage with SDG-relevant research. Your project can potentially complement this research, with a current specific focus on the geographical mapping of UZH publications and their alignment with the SDGs. Your work has the potential to offer valuable insights into how users might explore the geographic dimension of these contributions in the future.

UZH CO2 Emissions and Flight Data

- **Topic ID:** 13
- **Domain:** Sustainability and Environmental Impact in Academic Institutions
- **Dataset:** [**LINK**] available only from Thurs, October 25
- **Dataset Type:** multivariate (num, cat, mixed)

- **User group:** UZH Sustainability Team, UZH departments, general public, NGOs
- **Motivation/Goal:** In an era of increasing awareness about climate change and the urgent need for sustainability, it's essential for institutions like UZH to actively address their environmental impact. One significant contributor to carbon emissions is air travel, which is often a necessary part of academic and research activities. UZH recognizes this and has made efforts to monitor and mitigate its air travel-related CO2 emissions, as indicated by the information available on the UZH Sustainability page. The student project suggested here aims to take an explorative approach to comprehensively understand and address the CO2 emissions associated with air travel at UZH. We'll delve into the data, including essential factors like travel dates, departure and arrival locations (including layovers), flight class, etc. The dataset covers the years 2018 - 2022. Through the explorative approach, we aim to not only quantify emissions but also uncover patterns and opportunities for improvement; with a special emphasize on the geography. Furthermore, the insights generated by this project hold the potential to be integrated into an internal web application. Depending on the results and the usefulness of the data, this web application could provide a dynamic and interactive platform for the UZH community to explore and engage with the environmental impact of their air travel choices.

Soccer Players: Similarity, Search, and Exploration

- **Topic ID:** 14
- **Domain:** Sports. Every human is unique, and so is their notion and perception of item similarity. How similar are Lionel Messi and Christiano Ronaldo? Human-centered definitions of similarity will allow humans to execute meaningful nearest-neighbor searches, execute semantically meaningful clusterings, and benefit from other more human-centered ML techniques along these lines. An interesting case to study human-centered similarity is European soccer. Soccer receives a lot of attention from different stakeholders including clubs, managers, talent scouts, fans, TV industry, and marketing.
- **Dataset:** See Sharepoint. The recommendation is given to start with the European top five leagues (England, France, Germany, Spain, Italy), and possibly extend the considered dataset to all players later in the process.
- **Dataset Type:** tabular, multivariate (cat, num, mixed)
- **User group:** Fans, in particular, but possibly different stakeholders, each may receive their own similarity definition that is interactively selectable.
- **Motivation/Goal:** IVDA-supported research into item similarity enables people to calibrate interactive ML systems, with the effect that the similarity function used by the ML adapts towards user preferences. A quite promising field of human-in-the-loop research. In the soccer case, an IVDA tool is needed that provides multiple definitions of player similarity, allowing users to seamlessly switch between the different similarity definitions. The similarity definitions for players should enable users to formulate nearest neighbor searches (query-by-example, such as nearest neighbors for Lionel Messi), and explore all players in a dimensionality-reduced visualization of the player data, optionally, an interactive clustering can be added that detects patterns of soccer players. Ultimately, a most human-centered functionality would be to enable users define their similarity function at runtime, meaning that depending on the preferences of users, the similarity function used by the system adapts to humans (example: by humans describing the importance of player attributes).

Human-Centered Relation Discovery

- **Topic ID:** 15
- **Domain:** Finance
- **Dataset:** See Sharepoint
- **Dataset Type:** temporal, multivariate (num, cat, mixed)
- **User group:** Finance experts, Investors
- **Motivation/Goal:** Finding relations among attributes over time can help investors to better understand the behavior of stocks. One way to find relations among attributes is correlations. Another way to find relations is binning the numerical values to low, average and high based on the quantiles. Then, they can find which two attributes and their sub-ranges are related to each other. However, sometimes relations may not be between individual attributes but with combinations of attributes. For example, high dividend and high market cap together show a higher

correlation to the attribute returns. We want to find these relations and let a user to construct groups of attributes interactively to discover insightful relations.

Exploratory Analysis of Graph Data

- **Topic ID:** 16
- **Domain:** Interaction among people is often represented as a graph where nodes represent people and links denote relationship among them. For example, in communication networks, a link between two nodes may mean that node A called or sent a message to node B. Clustering helps to reveal patterns in the data such as close people to the person of interest in communication networks.
- **Dataset:** Stanford Large Network Dataset Collection
  - Amazon product co-purchasing network and ground-truth communities: `https://snap.stanford.edu/data/com-Amazon.html`
  - DBLP collaboration network and ground-truth communities: `https://snap.stanford.edu/data/com-DBLP.html`
- **Dataset Type:** Undirected graphs with their ground-truth communities. Each graph is a file of edge lists organized into two columns. Each row represents a link between two nodes. For example: "a b", "b a". The above edge list contains two nodes and an undirected link between them (from a to b and from b to a). Regarding the ground-trough communities (clusters), each row represents a community.
- **User group:** Data Analyst
- **Motivation/Goal:** Interactive integration of common clustering algorithms such as K-means, Louvain, Girvan–Newman algorithm and Label Propagation in a web-based Interactive Visual Data Analysis (IVDA) tool. The tool should allow a user to import a graph as a text file (edge list) and visualize the clusters generated by the selected algorithm. The tool should provide an option of importing ground-truth clusters (when available) for the imported graph. In this case, the tool should visualize both the clusters generated by an algorithm and the ground-truth clusters for comparison purposes. Different ways exist for layouts of graphs in 2D, ideally, users can switch between several layouts interactively. Preprocessing: Amazon and DBLP graphs are big and difficult to visualize as entire graphs. The preprocessing task involves sampling small sub-graphs of less than 1000 nodes that can be imported into the IVDA tool for visualization. The ground-truth communities contain duplicates, which can be removed in the pre-processing phase.