Project instructions
Data Mining I

The objective of the project is to test the knowledge and skills you've gained during the course through a real Data Mining process. This project is intentionally designed as an independent activity, where you are expected to identify relevant questions that can be answered using Data Mining methods, independently analyze the problems you encounter, and determine appropriate solutions. The project uses real data, without any simplifications made to highlight educational concepts, as was done in the preparatory assignments.

GROUPS

The project is completed by groups formed on Studium. However, grades are individual: each member must present a portion of the project work, and will be asked questions to verify their understanding of the underlying theory. For example, if a group member is responsible for identifying association rules, we might ask them to define *confidence*, explain when and why confidence can produce misleading results, and discuss alternative measures that could have been used.

PROJECT STRUCTURE AND GRADING

- The project is graded based on pass/fail (U/G) grades.
- You will be able to book an examination time for your group after the groups have been finalized on Studium.
- To pass the project, you must:
  1. Define a Data Mining **question/problem** based on the data you have chosen, with the following requirements:
     a. The question must be expressed in non-technical terms. For example, “Can we identify any discrimination between men and women based on census data?” Questions like “We’ll do clustering!” will not be accepted. A person without any knowledge of Data Mining should be able to understand what you want to do and why the question is relevant.
     b. You must explain why this question is important, such as how you would use the newly acquired knowledge.
     c. The question must require the application of at least one of the Data Mining algorithms studied in this course and cannot be answered using just SQL or basic descriptive statistics.

  2. Identify the **Data Mining algorithm(s)** that are appropriate for answering your question. For example, you might build an anomaly detection algorithm to detect fraud in credit card transactions, or use association rule mining to identify customer patterns in online shopping data.
  3. **Preprocess** the data appropriately to prepare it for the selected Data Mining algorithms.

4. Execute the chosen Data Mining algorithms and present the obtained results in an oral presentation. Since you are working with real data, it is possible that you may not identify any interesting **patterns**, which is acceptable; we evaluate the process, not the results. In fact, claiming to have found patterns that are only weakly supported by data and theory would be discouraged
5. **Interpret** your results and be prepared to **convince your examiner** that your findings are reliable, whether you have identified patterns or are claiming that no patterns exist in the data.
6. Demonstrate that you **know the theory** underlying your work.

To ensure transparency in the examination process, we provide detailed examination criteria as a checklist, available at the end of this document. To pass, you must satisfy all relevant criteria. If any criteria are not met, you may be asked to revise part of the project and/or deliver an additional presentation. Higher grades can be achieved based on the quality of your work and your performance during the oral examination.

- You must submit the slides for your presentation on Studium before the deadline indicated online. The presentation should include all the necessary information to demonstrate that the relevant items on the checklist have been addressed. The reason for submitting your slides is to ensure that all students have the same amount of preparation time.

- If you miss the deadline or do not pass the oral examination, you will have an additional opportunity to pass the project. After resubmission of your project, you will be contacted by the teacher to schedule a date and time for the presentation. Additional information, including exact dates for resubmissions and re-examinations, will be provided later.

TOOLS

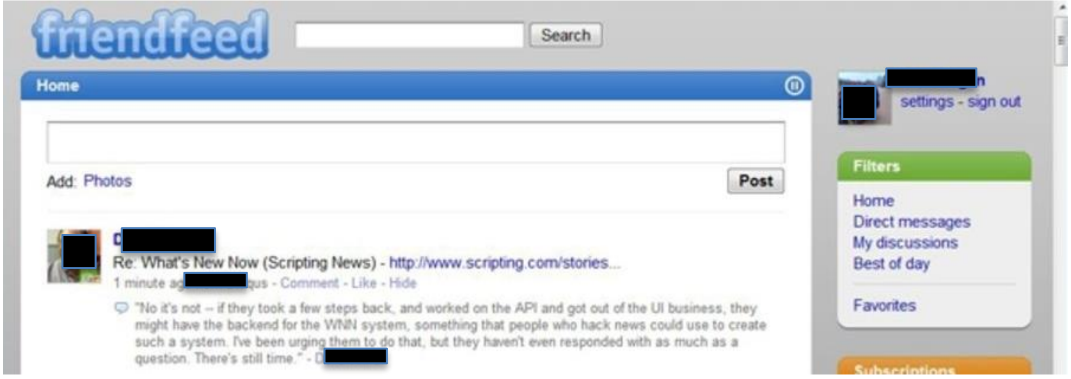
You are free to choose any tools to perform your analysis. A recommended combination is MySQL or any other relational DBMS if you need extensive preprocessing power and want to leverage your SQL knowledge, along with Python. However, you are free to choose other tools or languages if you prefer, such as R, MATLAB, etc. No approval is needed for your choice.

You are expected to bring your process/code to the presentation, as we may ask you to execute certain parts of it live, possibly changing parameters, etc.

DATA

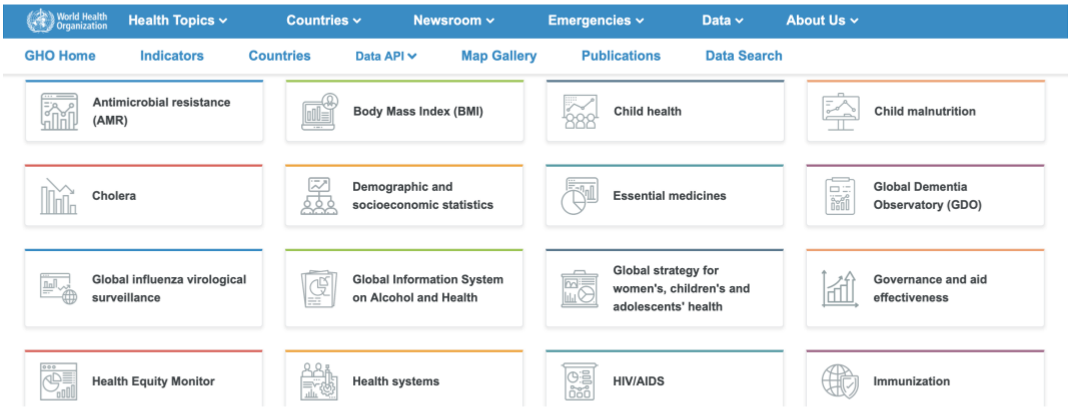
For your analysis, you can use a dataset of your choice and interest (that must be approved by your tutor) or one of the following data sources (that do not need approval):

- Data from the Friendfeed study, obtained by monitoring an Online Social Network, with user posts, likes, following/followers, etc.:
  https://drive.google.com/folderview?id=0B_D5tuT1vDQtckFGWkk1aTh5VlE&usp=sharing

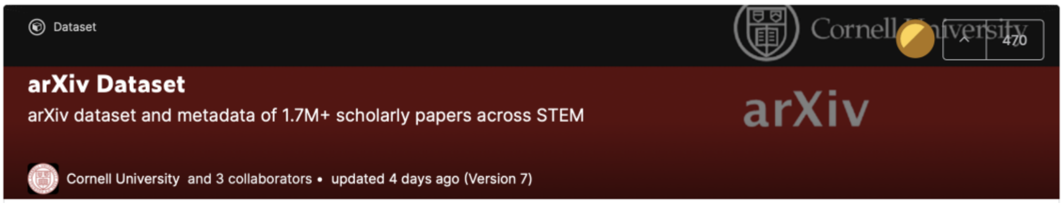


D[redacted]
Re: What's New Now (Scripting News) - http://www.scripting.com/stories...
1 minute ag[redacted]qus - Comment - Like - Hide
"No it's not -- if they took a few steps back, and worked on the API and got out of the UI business, they might have the backend for the WNN system, something that people who hack news could use to create such a system. I've been urging them to do that, but they haven't even responded with as much as a question. There's still time." - D[redacted]

- Data from the Global Health Observatory Data Repository: https://www.who.int/data/gho


World Health Organization
Health Topics
Countries
Newsroom
Emergencies
Data
About Us
GHO Home
Indicators
Countries
Data API
Map Gallery
Publications
Data Search
Antimicrobial resistance (AMR)
Body Mass Index (BMI)
Child health
Child malnutrition
Cholera
Demographic and socioeconomic statistics
Essential medicines
Global Dementia Observatory (GDO)
Global influenza virological surveillance
Global Information System on Alcohol and Health
Global strategy for women's, children's and adolescents' health
Governance and aid effectiveness
Health Equity Monitor
Health systems
HIV/AIDS
Immunization

- Data from the arXiv repository, with information about research papers in STEM: https://www.kaggle.com/Cornell-University/arxiv

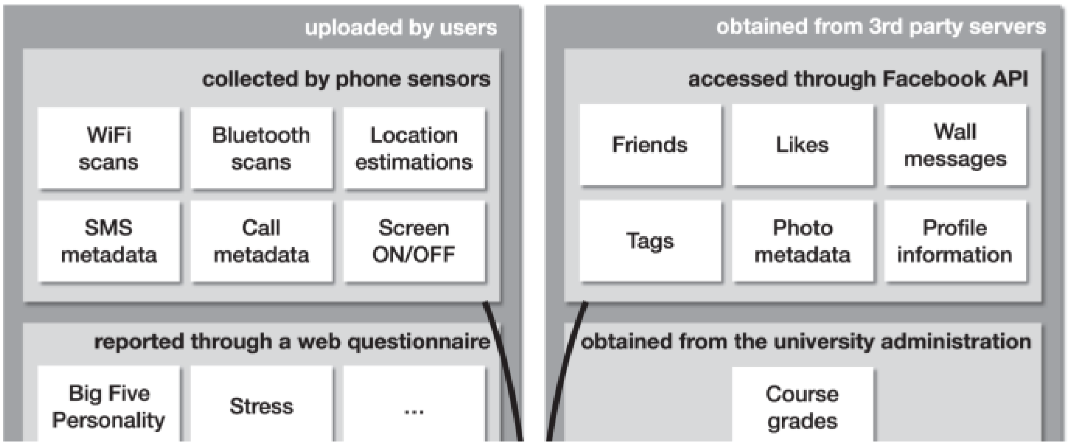


arXiv Dataset

arXiv dataset and metadata of 1.7M+ scholarly papers across STEM

Cornell University and 3 collaborators • updated 4 days ago (Version 7)

- Data from the Copenhagen Networks Study, with interactions between University students: https://www.nature.com/articles/s41597-019-0325-x



These data sources should give you an indication of what we expect if you choose your own data. In general, **we will not accept simple data sources that have already been prepared for specific analyses**, as is the case with many Kaggle datasets. Some Kaggle datasets (such as the one listed above) may still be acceptable, provided you choose an original question that **requires some preprocessing and non-trivial analysis**.

Regardless of the dataset you choose, keep in mind that **you will likely need to spend most of the project time understanding, retrieving, and preprocessing the data**. This is one of the intended learning outcomes of the project and cannot be fully achieved through lectures alone.

SUPPORT

This project tests your ability to independently design and execute a real-world knowledge discovery process– that has not been simplified for educational purposes or preprocessed to make patterns easy to discovery.

Therefore, you should work independently on this project.

However, we have planned two sessions where you will get feedback for your project work. Furthermore, you should feel free to contact the teaching assistants if you have further questions.

IMPORTANT DATES (DETAILS ON STUDIUM)

At the beginning of week 38, you must submit a one-page project proposal specifying the data you plan to use, your research question, and the rationale behind it. During week 38, we will either accept your proposal or request changes.

By the end of week 40, you must submit a preliminary report (in presentation format, e.g., PowerPoint) that includes a description of the initial data analysis, such as an overview of the main variables with basic visualizations like histograms and box plots, and the identification of any issues to be addressed, such as missing data or potentially incorrect values. The report should also detail your preprocessing techniques, model results, and a list of tasks remaining. You will receive feedback on your report during the project support session on Week 41, allowing you to make necessary changes before your final presentation.

Your final presentation deadline is October 15, 2025!

We hope you enjoy this experience, and we look forward to hearing your presentations!

CHECK LIST

☐ The problem is clearly stated at the beginning, in non-technical terms.

☐ The problem requires a Data Mining approach.

☐ The chosen Data Mining approach is appropriate.

☐ The chosen data representation (table, set, graph, …) is appropriate.

☐ [for tabular data] Each attribute has been given the correct type (nominal, …).

☐ The chosen algorithm is appropriate (motivate the choice w.r.t the features of the data and problem at hand).

☐ The list of applied pre-processing operations have been motivated.

☐ An appropriate proximity function (Jaccard, Manhattan, …) has been used, if required by the algorithm.

☐ If relevant, scaling/normalization issues have been addressed.

☐ If relevant, correlation issues have been addressed.

☐ Dimensionality has been reduced, if necessary.

☐ If relevant, appropriate train/test datasets have been generated.

☐ If necessary, noise has been reduced.

☐ If relevant, hyperparameters are tuned properly.

☐ The results are supported by sufficient evidence (large leaf sizes, high-support, …).

☐ The results have been interpreted, and related to the original problem (Was the problem solved? How can the results be used? Any new hypotheses have been generated?)