# Project Instructions

Data Mining I

Stephen Spengler, Georgios Panayiotou, Bahri Uzunoglu

## OVERVIEW

The objective of the project is to test the knowledge and skills acquired during the course on a real Data Mining process. This is, by design, an *independent* activity, where you are expected to identify relevant questions that can be answered using Data Mining methods, autonomously reason about the encountered problems and identify appropriate solutions. The project is performed on *real data*: nothing has been simplified to force some educational concepts to emerge, as done instead in the preparatory assignments.

## GROUPS

The project is performed by the groups formed on Studium. However, grades are individual: You will be able to acquire 60% of the points as a group, and the remaining 40% will be determined by your individual performance during the final oral presentation. For this, it is crucial that every single group member knows the whole project and can answer all questions.

## PROJECT STRUCTURE

- Choose a data source (see below) and define one Data Mining question / problem regarding the data you have chosen, with the following features:
    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 such as "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.
    b. You must be able to explain why that question is important and relevant, e.g., how you would be able to 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 performed e.g. just using SQL or basic descriptive statistics.
- Identify the Data Mining algorithms that are appropriate to answer your question. For example, setting up a classifier inspecting salaries, education, etc., or using association rule mining with filtered rules, …
- Preprocess the data appropriately, to make it ready for the selected Data Mining algorithms.
- Execute the chosen Data Mining algorithms and present the obtained results. Given that you work on real data, it may happen that you cannot identify any interesting patterns in the data – that would also be fine, we judge the process, not the results. In fact, claiming to have found patterns that are only weakly supported by data and theory would not be good.

- Interpret your results and be able to convince your examiner that your results can be trusted (whether you have found patterns, or you are claiming that there are no patterns in the data).

## SUBMISSIONS

Over the course of the project, you will have to submit the following documents:

1. A one-page project proposal, specifying the data you plan to use, your research question, and why you are asking that question. Additionally, you must give a (brief) explanation on how you will be using Data Mining to answer your research question.
2. A data exploration report, containing a description of the main variables with basic visualisations, such as histograms or box plots, and the identification of any issues to be addressed, such as missing data or potentially wrong values. This report can be done either as a text document or in a presentation format, but if you choose the latter, you might be able to reuse some of the slides for the final submission.
3. The final presentation. It should contain an introduction into your research question and a *brief* overview over the data set. The main part should focus on the algorithms that you have used and your results and conclusion.

All submissions must be done in pdf format on Studium before the deadlines stated there.

## EXAMINATION

Each submission will be graded and counts towards the final grade of the project. To make the examination transparent, we provide detailed examination criteria as a checklist.

1. Project Proposal [15 points]
   - [5] The research question is clearly stated at the beginning, in non-technical terms.
   - [5] Answering the research question requires a Data Mining approach (clustering, classification or association rule mining). The chosen Data Mining approach is appropriate.
   - [5] You explain why the research question is important and how you would be able to use the newly acquired knowledge.

<u>Note:</u> After submission and grading, your project coordinator will accept your proposal or request changes. In the latter case, you will need to resubmit, but the grade will stay the same. If the requested changes are not addressed in the resubmission, the project will be graded as failed.

2. Data Exploration [25 points]
   - [5] The chosen data representation (table, set, graph, …) is appropriate. Each attribute has been given the correct type.
   - [15] The data has been preprocessed. The applied preprocessing operations have been motivated.
   - [5] Computational dimensionality has been reduced.

3. Modelling [60 points]
    - [5] The chosen algorithm is appropriate (motivate the choice with respect to the features of the data).
    - [5] For the chosen algorithm, computational reductions, measures, approximations (e.g. appropriate proximity function, addressing correlation issues, …) are motivated.
    - [5] Modelling bias and errors (e.g. noise, overfitting, class imbalance) has been addressed.
    - [5] A good validation / evaluation test has been generated.
    - [5] The results are supported by sufficient evidence (large leaf sizes, high support, ...).
    - [5] The results have been interpreted, and related to the original problem (Was the problem solved? How can the results be used? Have any new hypotheses been generated?)

Note: This part of the project will be assessed in an oral clarification of the written project documents after the submission to assess the written project results. You will be able to book an examination time for your group after the groups have been finalised on Studium.

During the examination, each group member will individually answer exactly two of the above questions, as chosen by the examiner. Therefore, everyone needs to be prepared to answer each question.

The grading will be as follows. Each question is worth 5 points. For each question answered by one of your group colleagues, you will get their scored points, which amounts to a maximum of 20 points. For each question answered by yourself, you will get four times the score, for a maximum of 40 points. This totals to 60 points.

Note that if a student achieves less than the required individual grade to pass the last assignment, their scores will not be counted towards the final grade.

For example: Assume a group consisting of members A, B, C, each answering two questions

The final individual grade for each member will look like the table below:

| Question [ans. by, grade] | **A** | **B** | **C** |
|---|---|---|---|
| **Q1 [A, 4/5]** | 16 | 4 | 4 |
| **Q2 [B, 5/5]** | 5 | 20 | 5 |
| **Q3 [C, 3/5]** | 3 | 3 | 12 |
| **Q4 [A, 5/5]** | 20 | 5 | 5 |
| **Q5 [B, 2/5]** | 2 | 8 | 2 |
| **Q6 [C, 4/5]** | 4 | 4 | 16 |
| **Final individual grade (sum)** | 50 | 44 | 42 |

## GRADING DETAILS

The project is on the 5 / 4 / 3 / U grading scale:

| | |
|---|---|
| 5 [100-90) Pass with distinction: | Outstanding performance with only minor errors. |
| 4 [90-70) Pass with credit: | Generally sound work with a number of notable errors. |
| 3 [70-50) Pass: | Fair but with significant shortcomings. |
| U [50-0) Fail: | Fail – considerable further work is required. |

During the course, grading will be out of 100 internally and will be converted to Swedish System of U, 3, 4 or 5. Submissions are final, no changes are allowed after submission. If the task is delayed due to sickness or emergency, documentation needs to be provided. Otherwise project assignments submitted after the deadline will lose 20 points out of 100 for every day of late submission for 2 consecutive days. If the projects are submitted later than 2 days, grading will be done over 60 points out of 100 for that project.

Plagiarism is not allowed and suspected cheating leads to notification. Cheating will by default make you fail the assignment and/or exam and will cause more penalties.

If a student fails, a retake will be provided for the failed items that can make the student reach a passing grade for the late submission.

## TOOLS

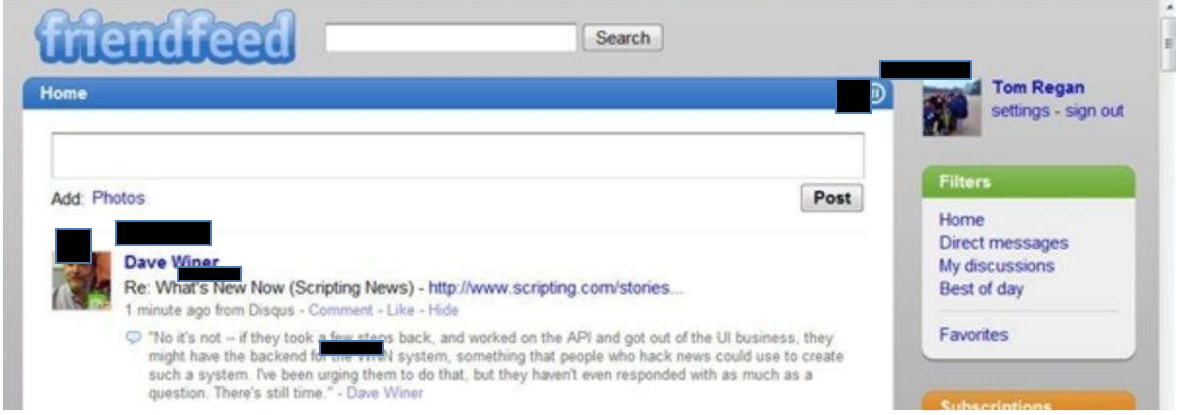
You are free to choose any tools to perform your analysis. A recommended combination is MySQL or any other relational DBMS (if you need a lot of initial preprocessing power and you want to exploit your knowledge of SQL) and RapidMiner, or only RapidMiner if you do not need database methods, but you are free to choose other tools/languages if you prefer, e.g., R, Python, etc. You do not need any approval for this.

You are expected to bring your process/code to the presentation, because we may ask you to execute some parts of it live, to change 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



[redacted]

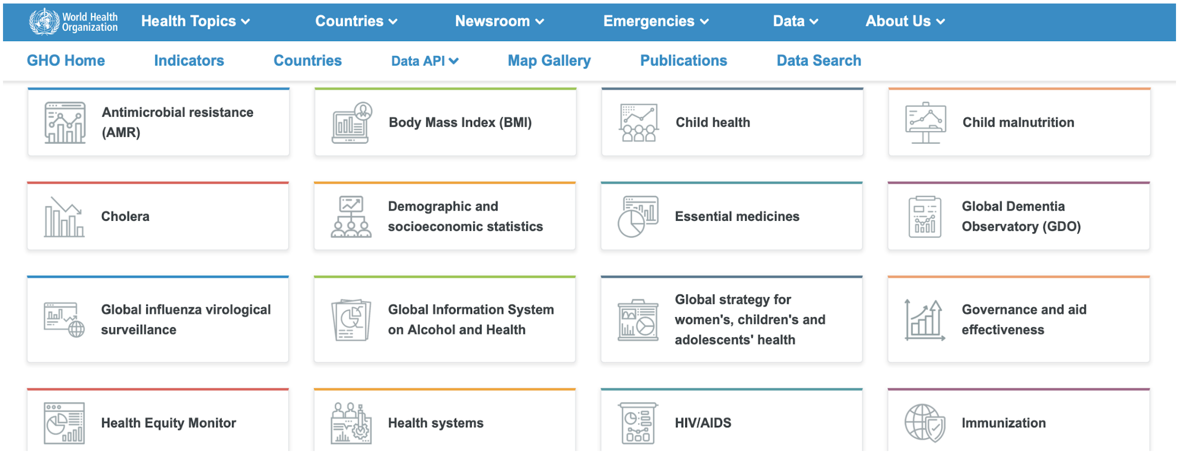
Dave Winer [redacted]

Re: What's New Now (Scripting News) - http://www.scripting.com/stories...

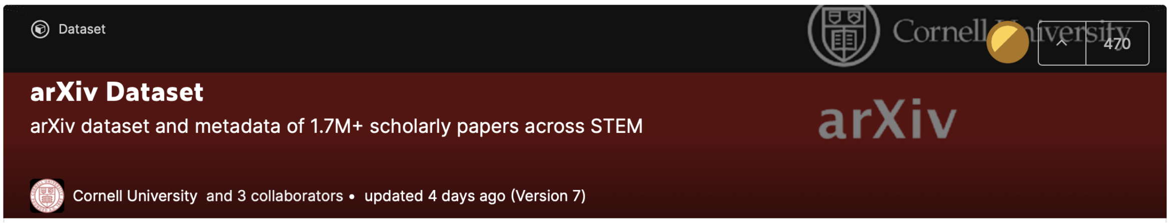
1 minute ago from Disqus - Comment - Like - Hide

"No it's not -- if they took [redacted] back, and worked on the API and got out of the UI business, they might have the backend for [redacted] 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." - Dave Winer

- 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

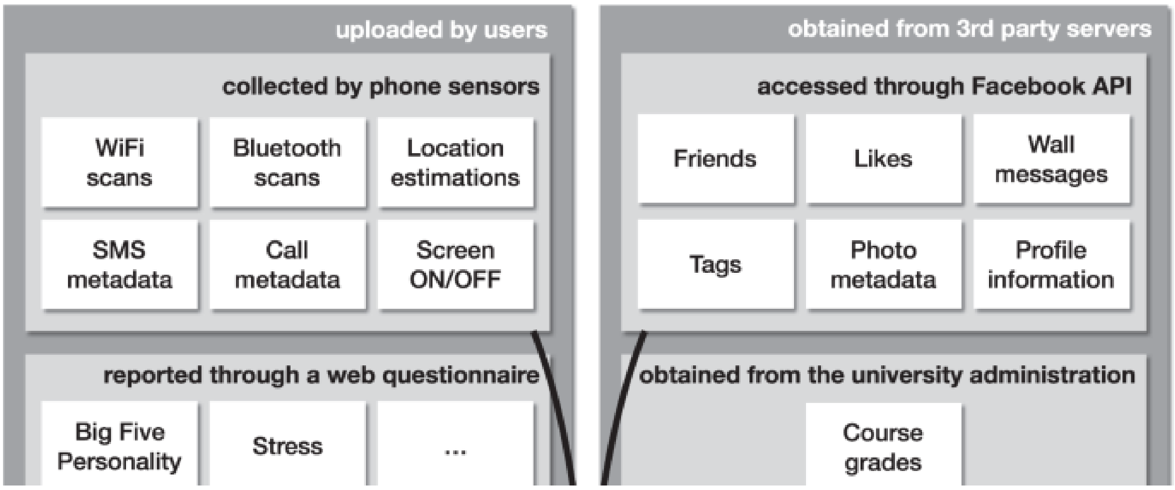


arXiv Dataset

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 a specific analysis, as is the case for many Kaggle datasets. Some Kaggle datasets (like the one listed above) can still be ok, if you choose an original question requiring some preprocessing and non-trivial analysis.

Independently of the chosen dataset, consider that you will 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, that cannot be obtained with the lectures.

**SUPPORT**

This project tests your ability to independently design and execute a knowledge discovery process on real data – that is, not "pedagogically" prepared to be simple to understand or where existing patterns have been "made ready for discovery".

Therefore, you should work independently on this project. Nevertheless, if you have further questions, feel free to contact your project coordinator at any time.

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