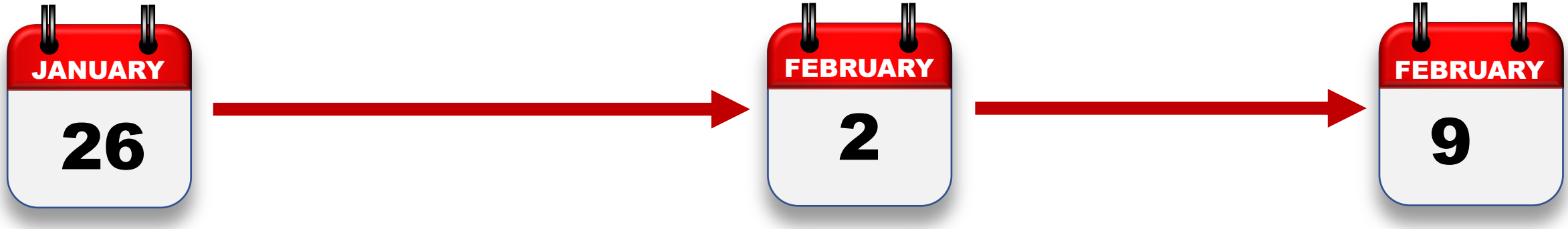
THE UNIVERSITY of EDINBURGH
informatics

Week 2: MSc Project Selection

THE UNIVERSITY of EDINBURGH
informatics



Jan 26th - Project Selection Period Opens

Browse the projects and contact potential supervisors

Feb 2 – All options contacted

Finalise selections and submit preferences

HARD DEADLINE FEB 9 *(dates for 2023/2034 session)*

IPP Week 2

- *Introduction to IPP was covered in Week 1 Sessions (Aurora Constantin)*

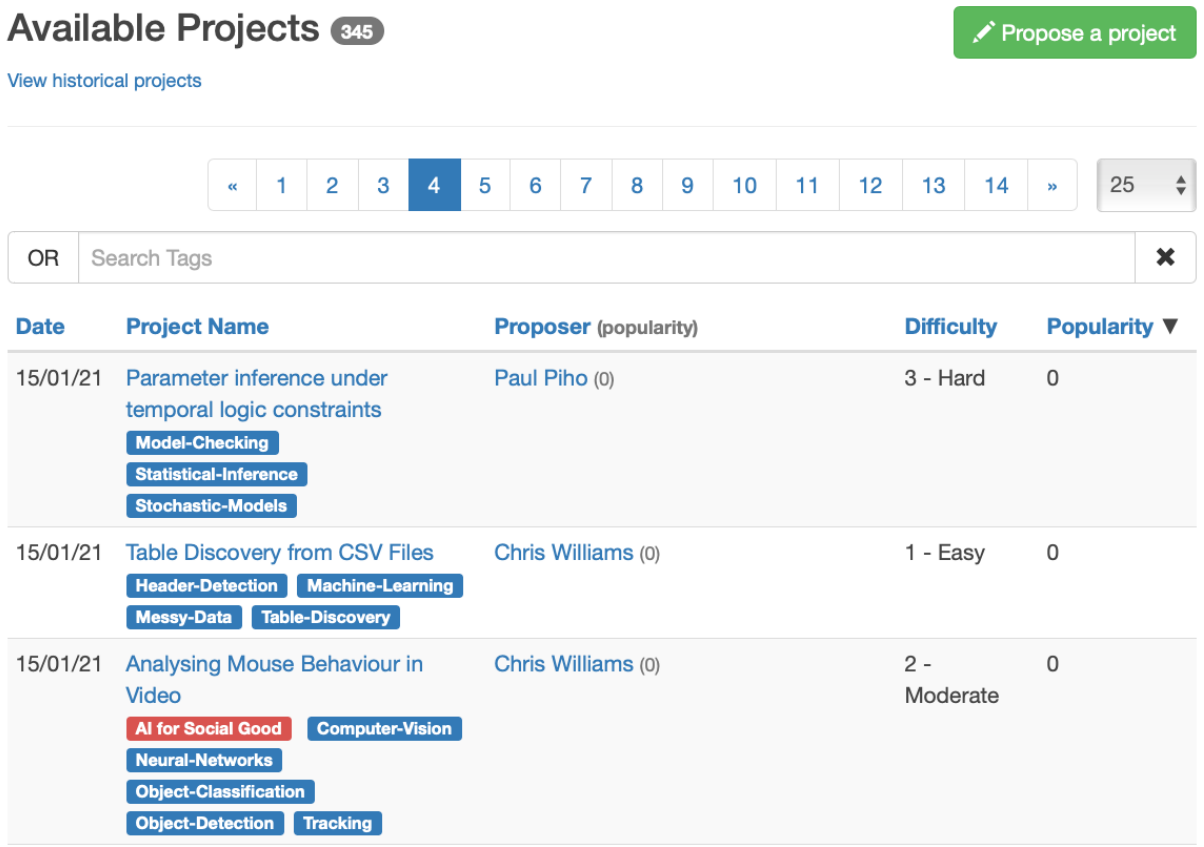
- MSc Project Selection
  - Process
  - Guidance
  - Allocation
  - FAQs
- *Open Q&A on Wednesday and on Piazza forum*

MSc Project Selection Process

Process in a nutshell…

- https://dpmt.inf.ed.ac.uk ***** ON CAMPUS *OR* VPN *****
- There you will see a (long) list of project titles (officially from Jan 26th)
- Each has a description and a proposer who will usually supervise the project.
- You express interest in the project (before Feb 2rd).
- The proposers assess your suitability (perhaps via a meeting, email exchange or some other method).
- You select projects you are interested in and rank these in order of preference (Feb 9 deadline).
- You can only select a project if the supervisor says you are suitable.
- DPMT allocates projects on the basis of suitability and preference.
- Ensures you are suitable for the project and tries to ensure everyone is "happy".

DPMT.inf.ed.ac.uk : Project List



THE UNIVERSITY of EDINBURGH
informatics

DPMT: Project Description

Table Discovery from CSV Files

Principal Goal: The goal of this research project is to evaluate algorithms for the detection of tables and headers in CSV files, and develop a new method based on CleverCSV [3].

This project has been proposed by Chris Williams (ckiw@inf.ed.ac.uk).

Plain-text data files are commonly used to store and transmit tabular data on the web. The comma-separated values (CSV) format is one of the most common file formats used for this purpose. However, the CSV format has no formal specification, which allows users to include additional data in the file other than the table itself, or include multiple tables in a single file. Because of this, manual intervention and cleaning is frequently needed before the data can be loaded.

Automatically extracting the tabular data from a given CSV file requires several steps: finding the location of the table(s) in the file, detecting the dialect of the CSV file (delimiter, quote character, etc.), and identifying where the header row is located (if present). There are some current solutions to this problem, e.g. [1], [2].

As part of two MSc projects last year, several hundred cvs files were annotated with their table and header regions, and this data is available. We also developed code to evaluate the quality of the results obtained from table discovery programs. In this project we will first evaluate [1] this dataset. Secondly, we will look to extend the CleverCSV program [3] to detect tables and headers, as suggested by Gerrit van der Burg (project co-supervisor) as per https://github.com/alan-turing-institute/CleverCSV/issues/29#issuecomment-724336107 . CleverCSV returns a summary of the row patterns for each row of the file, and it is likely that this can be used to identify table/header regions. A possible additional goal is to evaluate [2] on the dataset, but initial exploration of the code suggests it is incomplete, and it may be necessary to follow up with the authors of [2].

When contacting me to express interest in one of my projects, please send me information on courses taken and marks/grades obtained on your MSc. Please wait until you have any official semester 1 course marks for courses which were examined in December 2020.

1. T Doehmen, H Muehleisen, P Boncz. Multi-Hypothesis CSV parsing (2017). Proc 29th International Conference on Scientific and Statistical Database Management (SSDBM'17). [code available]
2. Pytheas: Pattern-based Table Discovery in CSV Files. C. Christodoulakis et al, Proc VLDB 13(11) 2075-2089 (2020). http://www.vldb.org/pvldb/vol13/p2075-christodoulakis.pdf [code available]
3. G. J. J. van den Burg, A. Nazabal, C. Sutton. Wrangling Messy CSV Files by Detecting Row and Type Patterns (2019). Data Mining and Knowledge Discovery, 33(6) 1799-1820. https://gertjanvandenburg.com/papers/VandenBurg_Nazabal_Sutton_-_Wrangling_Messy_CSV_Files_by_Detecting_Row_and_Type_Patterns_2019.pdf

DPMT: Project Description

Completion Criteria

Evaluation of [1] on the given datasets. Extension of [3] to address table discovery. Possible additional goal of evaluation of [2].

Additional Information

Difficulty
1 - Easy

Desirable Skills
Knowledge of programming in Python is preferred. Knowledge of R is desirable when comparing with previous work (i.e. [1]) Knowledge of working with the command line (or a willingness to learn this) is desirable.

Essential Skills
Basic knowledge of machine learning: performance of 60% or more on MLPR or IAML is required.

Resources Required
Standard computational resources.

Potential Supervisors

Supervisor		Email Address
Main	Chris Williams	ckiw@inf.ed.ac.uk

Good Practice

- **Spread** your project choices widely
  - Different fields (ML, NLP, Systems, …)
  - Different supervisors
  - Include less popular projects as backup
- **Read** the project descriptions thoroughly
- **Reach out** to project supervisors and be prepared!
  - Submit any **Questions** you have about the project?
  - Provide **Evidence** that you have the required background/skills!
- **Provide** reasons why you are **suitable** for the project

Bad Practice

- Select only very **popular projects**
- Select only projects from a **single supervisor**
- Select only projects from very **popular supervisors**
- Select a project **without sufficient background**
- Select only "**easy**" projects
- **Waste supervisors' time by:**
  - Not reading the project description
  - Not meeting the prerequisites
  - Being unprepared for the meeting with supervisor

Project Difficulty

- What does it **mean**?
  - easy, moderate, hard, very hard, challenging, variable
- What are "**easy**" projects?
  - May require fewer/no specific prerequisites
  - May have a lower initial passing hurdle
- Are project **grades capped** for "easy" projects?
  - No, but you need to go beyond the completion criteria to get high grades
  - Requires novel additions (beyond specification) for higher marks
- What does "**variable**" mean?
  - Project has scope for adjustment (stages, objectives)

Machine Learning/Natural Language Processing

- **Very high demand** for ML/NLP projects
- Edinburgh ML/NLP research groups are large, but …
  - … each supervisor's project supervision capacity is **limited**
- Most ML/NLP projects are likely to be extremely popular
- Not everyone who selects ML/NLP projects will get one
  - Don't select ML/NLP projects only!
  - Keep less popular projects as a **backup**

Project Allocation

Project Allocation

- Allocation is done **algorithmically**
  - Supervisors have no influence on this process
- Supervisors can mark you as
  - **not suitable, suitable, highly suitable**
- You must **select 5 projects** for which you have been marked as suitable!
- Many projects have **capacity** for >1 students

Self-proposed Projects (deadline has passed)

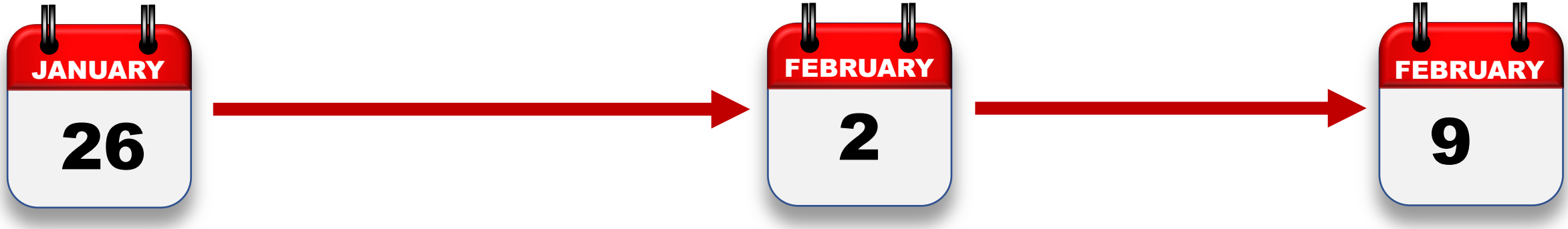
- Self-proposing a project is **not enough**
  - Project must also be **endorsed** by a supervisor
  - It's **your responsibility** to find a supervisor
  - If you can't find a supervisor, project will not go ahead
- Given **priority** in allocation, but …
  - … **no guarantee** that you will be allocated your self-proposed project
- Select 4 regular projects as **backup**

MSc Project Guide

- Authoritative Guide on everything related to MSc Project
- Compiled by MSc Project Manager
  - Amir Vaxman, msc-project-mgr@inf.ed.ac.uk
- https://opencourse.inf.ed.ac.uk/diss/

On Wednesday…

THE UNIVERSITY of EDINBURGH
informatics



Jan 26th - Project Selection Period Opens

Browse the projects and contact potential supervisors

Feb 2 – All options contacted

Finalise selections and submit preferences

HARD DEADLINE FEB 9