Homework 1

Special Directions on Machine Learning (SDML)
Fall 2019, NTU CSIE
Prof. Shou-De Lin
TA: Ywuan-Chai Chong

AI CUP 2019

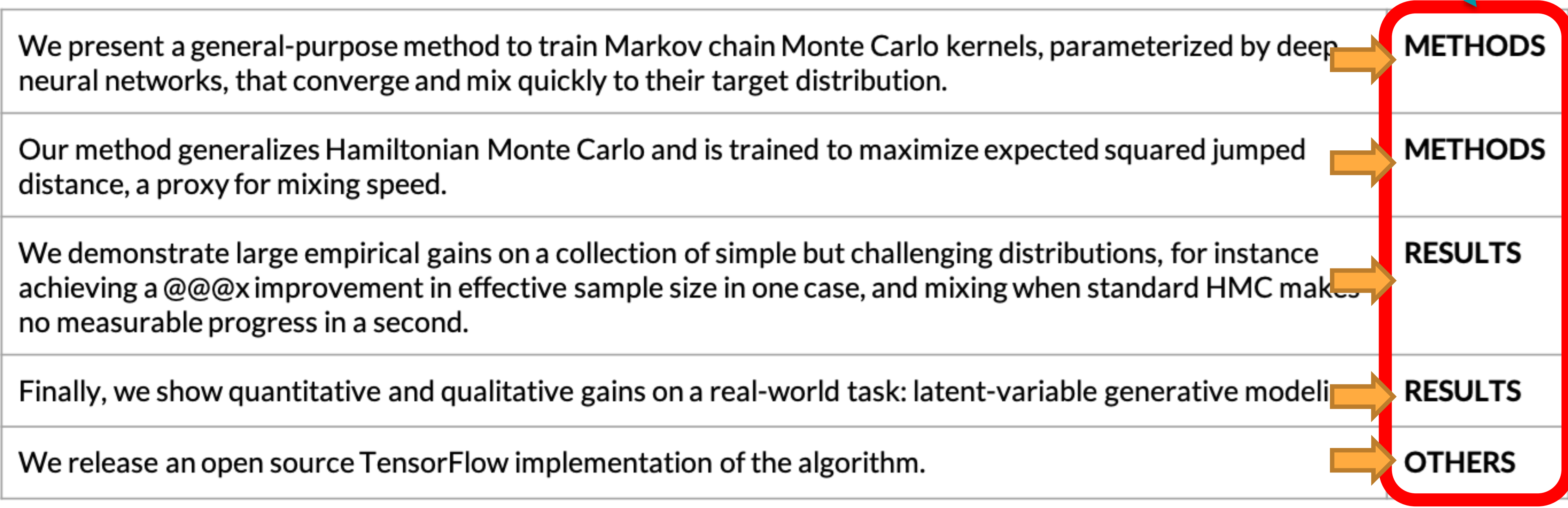
1. Abstract labeling 論文標註
   The contestants should use the provided materials to predict if a sentence in a thesis should be classified as the following categories: **Background**, **Objectives**, **Methods**, **Results**, **Conclusions**, or **Others**. Note that a sentence may have multiple classifications, e.g. a sentence may be classified as both Objective and Methods.

2. Abstract classification 論文分類 **[SDML HW1]**
   The contestants should use the provided materials to predict the classification of a thesis into the following categories: **Theoretical Paper**, **Engineering Paper**, **Empirical Paper** or **Others**. Note that a thesis may have multiple classifications, e.g. a thesis may be both a Theoretical Paper and an Engineering Paper.

AI CUP 2019 (cont.)

Abstract labeling

Paper Title: Generalizing Hamiltonian Monte Carlo with Neural Networks

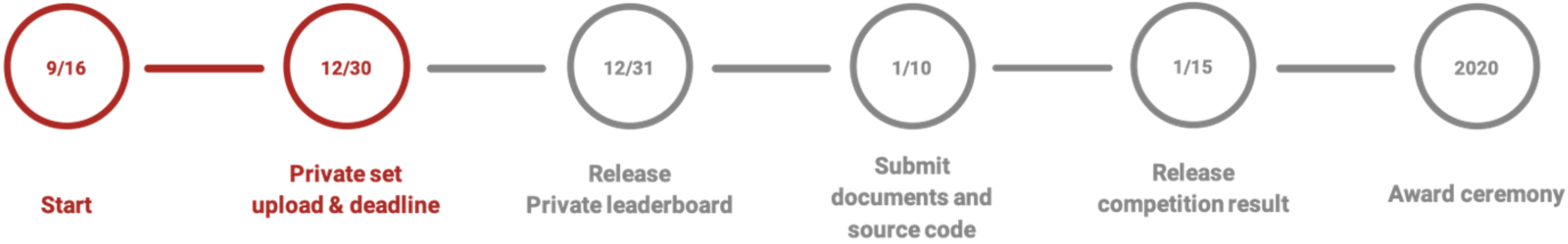


We present a general-purpose method to train Markov chain Monte Carlo kernels, parameterized by deep neural networks, that converge and mix quickly to their target distribution.	METHODS
Our method generalizes Hamiltonian Monte Carlo and is trained to maximize expected squared jumped distance, a proxy for mixing speed.	METHODS
We demonstrate large empirical gains on a collection of simple but challenging distributions, for instance achieving a @@@x improvement in effective sample size in one case, and mixing when standard HMC makes no measurable progress in a second.	RESULTS
Finally, we show quantitative and qualitative gains on a real-world task: latent-variable generative modeli	RESULTS
We release an open source TensorFlow implementation of the algorithm.	OTHERS

Paper types: **Theoretical**, **Empirical**

Abstract classification

AI CUP 2019 (cont.)



9/16 Start

12/30 Private set upload & deadline

12/31 Release Private leaderboard

1/10 Submit documents and source code

1/15 Release competition result

2020 Award ceremony

AI CUP 2019 (cont.)

1. Abstract labeling prizes (total 350,000 NTD):
   a. Trend Micro Elite Award in Artificial Intelligence: 100,000 NTD
   b. Student Group Leaderboards Champion: 100,000 NTD
   c. Student Group Leaderboards 2nd Runner Up: 60,000 NTD
   d. Student Group Leaderboards 3rd Runner Up: 40,000 NTD
   e. Student Group Leaderboards 4th ~ 9th place: 10,000 NTD
2. Abstract classification prizes (total 350,000 NTD):
   a. Trend Micro Elite Award in Artificial Intelligence: 100,000 NTD
   b. Student Group Leaderboards Champion: 100,000 NTD
   c. Student Group Leaderboards 2nd Runner Up: 60,000 NTD
   d. Student Group Leaderboards 3rd Runner Up: 40,000 NTD
   e. Student Group Leaderboards 4th ~ 9th place: 10,000 NTD

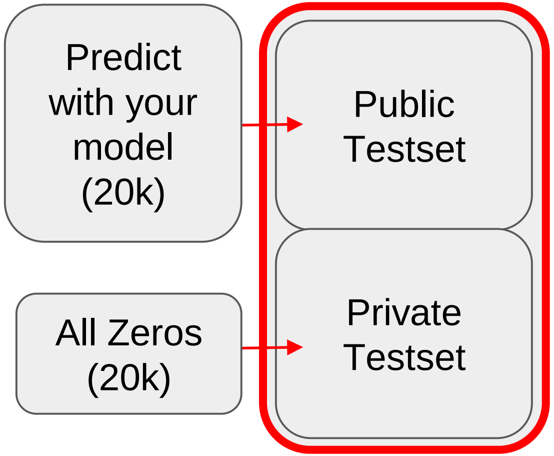
In addition to the top 8 ranking award for the Student Leaderboard Group, teams ranking in the top 25% and whose private score is above Baseline (Baseline: 0.69) will be presented with a Certificate of Workshop Award.

Introduction

- Topic: Abstract classification problem from **AI CUP 2019**.
- Data: Real-world research thesis (abstract).
- Extra Data: Real-world citation networks.

Abstract Classification

- Goal: To predict the classification of a thesis into the following categories:
  - Theoretical Paper
  - Engineering Paper
  - Empirical Paper
  - Others
  - **Note that a thesis may have multiple classification, e.g., a thesis may be both a Theoretical Paper and an Engineering Paper.**
- For the task of HW1
  - Real-world datasets are used. (Provided by AI CUP 2019)
  - All the labels are annotated by the authors of the paper.
  - Trainset 7000
  - Public Testset 20000
  - Private Testset 20000
  - Sample Submission 40000



Data

Id 流水號

Title 論文題目

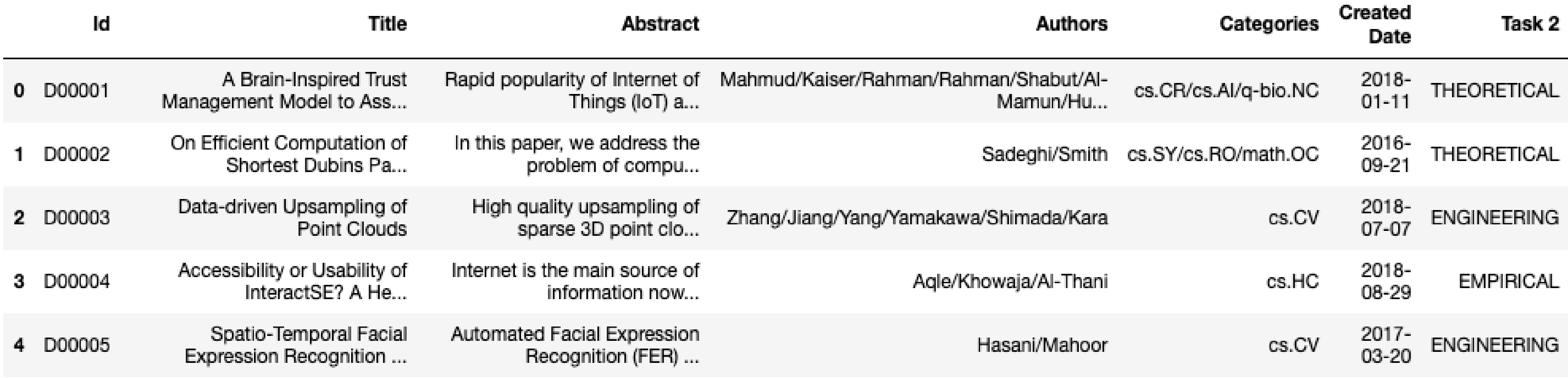
Abstract 論文摘要

Authors 作者們

Categories 論文類別

Dates 發佈日期

Label 分類類別



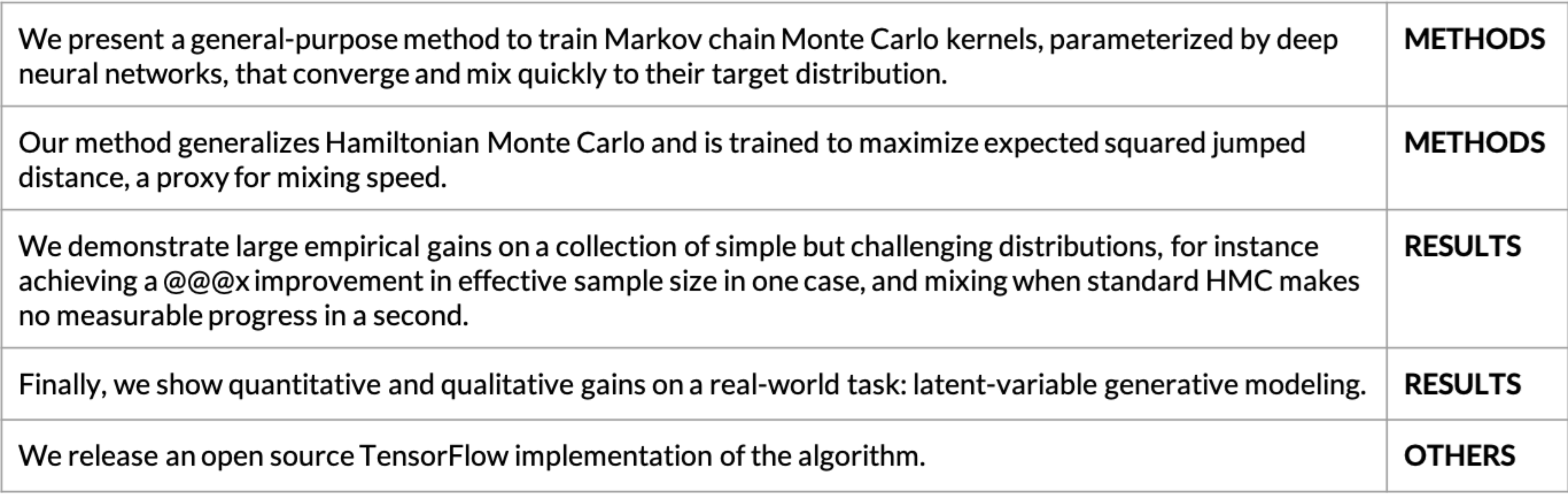
	Id	Title	Abstract	Authors	Categories	Created Date	Task 2
0	D00001	A Brain-Inspired Trust Management Model to Ass...	Rapid popularity of Internet of Things (IoT) a...	Mahmud/Kaiser/Rahman/Rahman/Shabut/Al-Mamun/Hu...	cs.CR/cs.AI/q-bio.NC	2018-01-11	THEORETICAL
1	D00002	On Efficient Computation of Shortest Dubins Pa...	In this paper, we address the problem of compu...	Sadeghi/Smith	cs.SY/cs.RO/math.OC	2016-09-21	THEORETICAL
2	D00003	Data-driven Upsampling of Point Clouds	High quality upsampling of sparse 3D point clo...	Zhang/Jiang/Yang/Yamakawa/Shimada/Kara	cs.CV	2018-07-07	ENGINEERING
3	D00004	Accessibility or Usability of InteractSE? A He...	Internet is the main source of information now...	Aqle/Khowaja/Al-Thani	cs.HC	2018-08-29	EMPIRICAL
4	D00005	Spatio-Temporal Facial Expression Recognition ...	Automated Facial Expression Recognition (FER) ...	Hasani/Mahoor	cs.CV	2017-03-20	ENGINEERING

Data (cont.)

- Id
  - Serial number.（流水編號，無特別意義。）
- Title
  - Title of the paper.（論文文章標題。）
- Abstract
  - Abstract of the paper. Each sentence is be separated by $$$. （論文摘要，以 $$$ 將句子隔開。）
- Authors
  - Authors of the paper. Each name is be separated by /.（論文作者，每個作者以 / 將句子分開。）
- Categories
  - Each field is be separated by /.（該論文在arXiv上的分類，多個分類以 / 做分割。）
- Created Date
  - The date which the paper uploaded to www.arxiv.com.（論文上傳至www.arxiv.com 的日期。）
- Task 1 / Task 2 Label
  - T1 - Multilabel of each sentence in the paper. Each multilabel is be separated by space and each label is be separated by /.（ 論文摘要的句子分類，每個句子的分類以 空格 分開，同個句子多個分類以 / 分開。 ）
  - T2 - Multilabel of the paper. Each label is be separated by space.（ 論文的分類，多個分類以 空格 做分割。 ）

Data (cont.)

Generalizing Hamiltonian Monte Carlo with Neural Networks

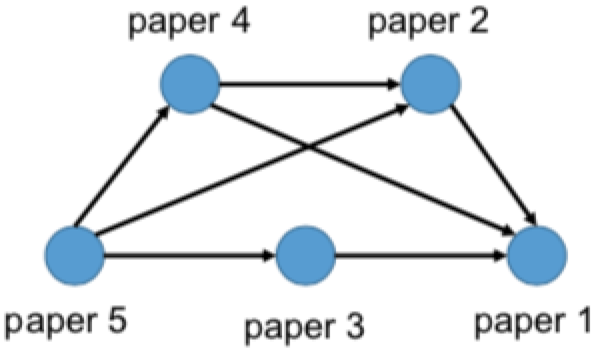


We present a general-purpose method to train Markov chain Monte Carlo kernels, parameterized by deep neural networks, that converge and mix quickly to their target distribution.	**METHODS**
Our method generalizes Hamiltonian Monte Carlo and is trained to maximize expected squared jumped distance, a proxy for mixing speed.	**METHODS**
We demonstrate large empirical gains on a collection of simple but challenging distributions, for instance achieving a @@@x improvement in effective sample size in one case, and mixing when standard HMC makes no measurable progress in a second.	**RESULTS**
Finally, we show quantitative and qualitative gains on a real-world task: latent-variable generative modeling.	**RESULTS**
We release an open source TensorFlow implementation of the algorithm.	**OTHERS**

Paper types: **Theoretical**, **Empirical**

Extra Data - Citation Networks

- Node: publications (e.g., papers, articles, books)
- Edge: citation references
  - Directed edges (arcs) <a,b> =/= <b,a>
- Directed Acyclic graphs (DAGs)
  - Temporal topological order on citations (i.e., links).



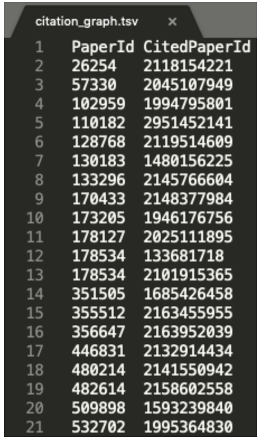
⟺

paper 5 cites paper 4;
paper 5 cites paper 3;
paper 5 cites paper 2;
paper 4 cites paper 2;
paper 4 cites paper 1;
paper 3 cites paper 1;
paper 2 cites paper 1

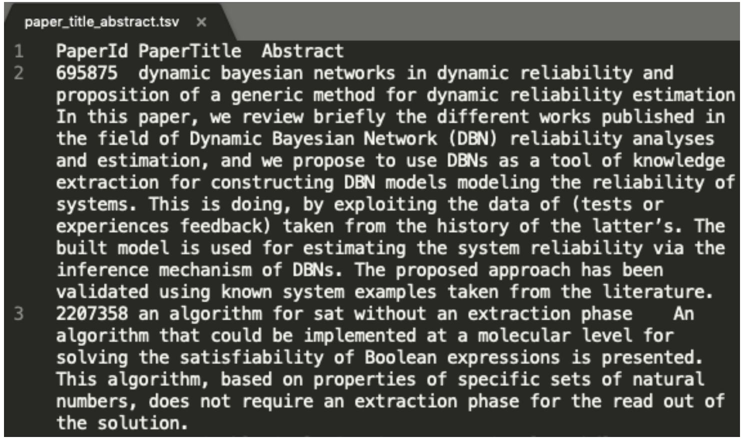
Extra Data - Citation Networks (cont.)

- mag_paper_data/

  |-- citation_graph.tsv
  |-- paper_title_abstract.tsv
  |-- README
  |-- others.
- This graph contains
  - ~770k nodes, ~1372k edges
  - ~708k nodes have Title and Abstract
  - 1780 connected components,
  - Largest connected components included 760k nodes, 98.77% nodes are linked



```
PaperId CitedPaperId
26254   2118154221
57330   2045107949
102959  1994795801
110182  2951452141
128768  2119514609
130183  1480156225
133296  2145766604
170433  2148377984
173205  1946176756
178127  2025111895
178534  133681718
178534  2101915365
351505  1685426458
355512  2163455955
356647  2163952039
446831  2132914434
480214  2141550942
482614  2158602558
509898  1593239840
532702  1995364830
```

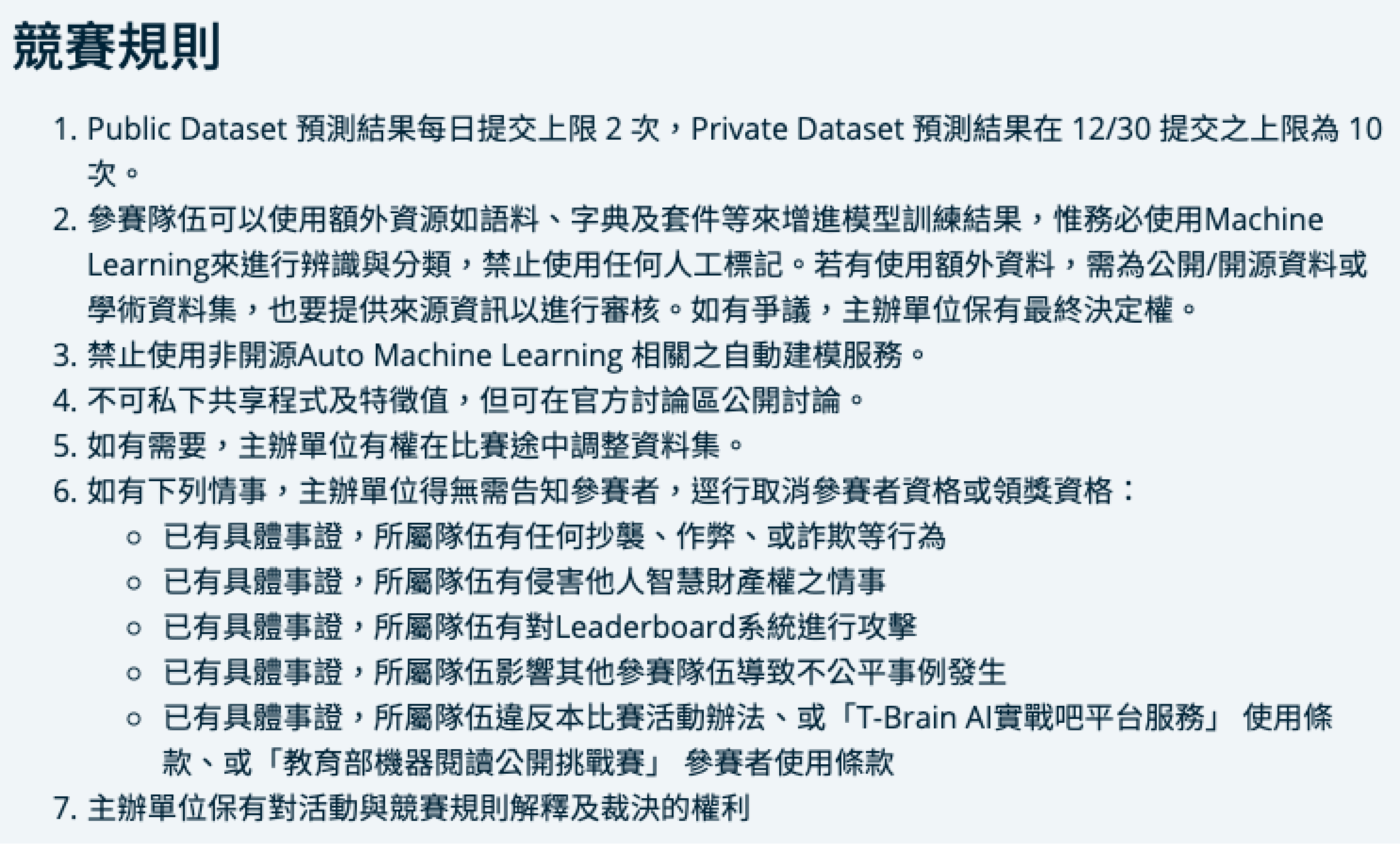


```
PaperId PaperTitle  Abstract
695875  dynamic bayesian networks in dynamic reliability and
proposition of a generic method for dynamic reliability estimation
In this paper, we review briefly the different works published in
the field of Dynamic Bayesian Network (DBN) reliability analyses
and estimation, and we propose to use DBNs as a tool of knowledge
extraction for constructing DBN models modeling the reliability of
systems. This is doing, by exploiting the data of (tests or
experiences feedback) taken from the history of the latter's. The
built model is used for estimating the system reliability via the
inference mechanism of DBNs. The proposed approach has been
validated using known system examples taken from the literature.
2207358 an algorithm for sat without an extraction phase    An
algorithm that could be implemented at a molecular level for
solving the satisfiability of Boolean expressions is presented.
This algorithm, based on properties of specific sets of natural
numbers, does not require an extraction phase for the read out of
the solution.
```

- You may find some papers from Trainset (7k) + Public Testset (20k) + Private Testset (20k) which do not appear in this graph.
- Using thesis title to match between dataset (trainset + testset) with this graph.
  - Please take a look on README - “How to normalize thesis title for mapping.”
- Download link

TBrain Competition

- Testing data will be divided into public and private testing.
  - You will be evaluated based on the performance of **public testing** only.
- Maximum 2 submissions a day are premitted.
  - **Last valid submission score** will be choosen as final score before the deadline.
  - Remember to declare your team on the Tbrain platform.
- Using *extra data* from the Internet is prohibited.
- All rules and regulations follow AI CUP 2019.
- Please use the SDML_<Student-ID> or SDML_<Group Name> as the TBrain nickname to show on the leaderboard.
  - Please form a team (not more than 3 people) and fill in HERE (google sheets).
  - For example, SDML_r07922001 or SDML_Team01 as the nickname
- Competition pages:
  - https://tbrain.trendmicro.com.tw/Competitions/Details/9



競賽規則

1. Public Dataset 預測結果每日提交上限 2 次，Private Dataset 預測結果在 12/30 提交之上限為 10 次。
2. 參賽隊伍可以使用額外資源如語料、字典及套件等來增進模型訓練結果，惟務必使用Machine Learning來進行辨識與分類，禁止使用任何人工標記。若有使用額外資料，需為公開/開源資料或學術資料集，也要提供來源資訊以進行審核。如有爭議，主辦單位保有最終決定權。
3. 禁止使用非開源Auto Machine Learning 相關之自動建模服務。
4. 不可私下共享程式及特徵值，但可在官方討論區公開討論。
5. 如有需要，主辦單位有權在比賽途中調整資料集。
6. 如有下列情事，主辦單位得無需告知參賽者，逕行取消參賽者資格或領獎資格：
   - 已有具體事證，所屬隊伍有任何抄襲、作弊、或詐欺等行為
   - 已有具體事證，所屬隊伍有侵害他人智慧財產權之情事
   - 已有具體事證，所屬隊伍有對Leaderboard系統進行攻擊
   - 已有具體事證，所屬隊伍影響其他參賽隊伍導致不公平事例發生
   - 已有具體事證，所屬隊伍違反本比賽活動辦法、或「T-Brain AI實戰吧平台服務」 使用條款、或「教育部機器閱讀公開挑戰賽」 參賽者使用條款
7. 主辦單位保有對活動與競賽規則解釋及裁決的權利

Grading Policy

- Performance (50%)
    - *Performance Ranking:* all the participant will be ranked according to the TBrain Public testing scores.
    - *Baselines:* you need to beat our baseline for a basic score.
        - Simple GRU tutorial (0.68)
        - Bert: 0.716
        - **BASELINE: 0.7**
- Report (50%)
    - *Coverage* (25%): #methods you tried; please describe and analyze the approaches with experiment results.
    - *Novelty* (25%): how novel is your model designed. (Ensemble techniques are valid, however we encourage novel single models.)

CEIBA Submissions

- You should submit your score codes along with reports to the corresponding CEIBA entries.
  - Including any third-party scource codes you used.
  - A .zip file should be uploaded.
  - The format of CEIBA submissions is stated later.
- Your CEIBA submissions should match your final output on the TBrain platform.
  - That is, your source codes should be able to reproduce your final performance scores on TBrain.
- Plagiarism is strictly prohibited.
  - You should clearly mention all the third-party codes (if any) used in your submissions.
  - We will check your source codes via professional softwares.

Reports

- Your report should be formatted in PDF files.
- Only digital submissions on CEIBA are acceptable.
- The report should include:
  - Official name and the student ID of each member.
  - Attempted approaches to solve specific problems.
  - Analyses and observations based on experiment results.
  - Difficulties encountered, unsolved issues, etc.
- No page limit. ☺
  - Feel free to include all the experiment results, reference theorems or other appendices.

Format of CEIBA Submissions

```
[student-id].zip (team leader’s, e.g., r07922001.zip)
|-- src/ (the source codes written by you)
|-- lib/ (all the libraries, third-party source codes you used)
|-- report.pdf
|-- README (a ‘plaintext’ file to explain how to reproduce your results)
```

(You *must* submit this .zip to get the 50% performance points.)

<Important> You should upload in **.zip** format.

.rar, .tar, .gz, .7z, or any other formats will receive **0 points** without grading.

Submission Deadlines

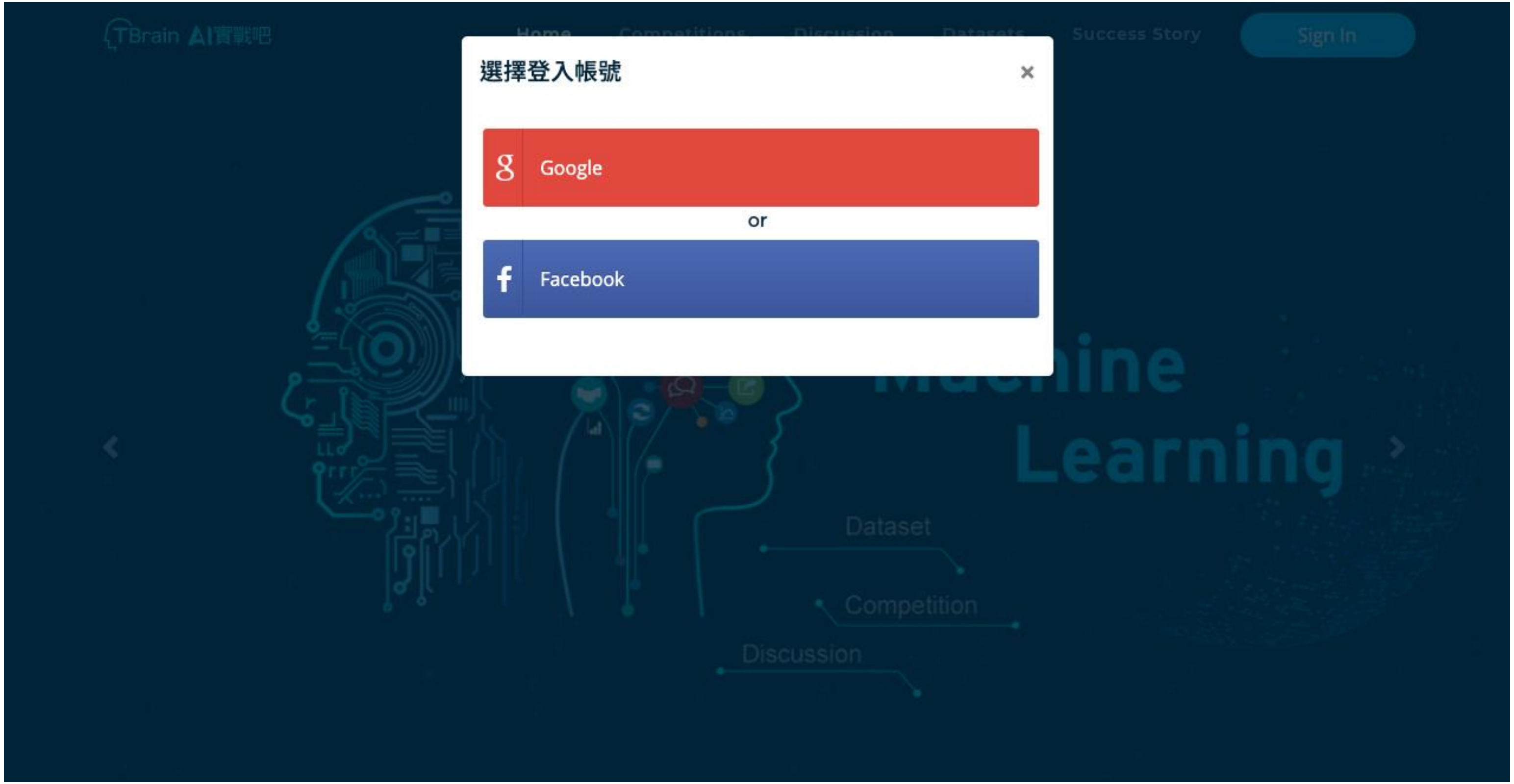
- Due time: 2019/10/15 (Tue.) 23:59:59 (Taiwan time)
  - According to the system times of TBrain and CEIBA
  - Since the network status is unpredictable, please make your submissions as earlier as possible.
  - Only team leader has to submis the .zip & report.
  - Report due on: 2019/10/20 (Sun.) 23:59:59
- HW1 presentation: 10/17 (Thur.)
  - Upload your presentation slide to CEIBA before 13:00:00 (Taiwan time).
- For the delayed submissions:
  - Within 24 hours: original_task_score * 0.5
  - More than 24 hours: zero point for the homework.

Contact TA

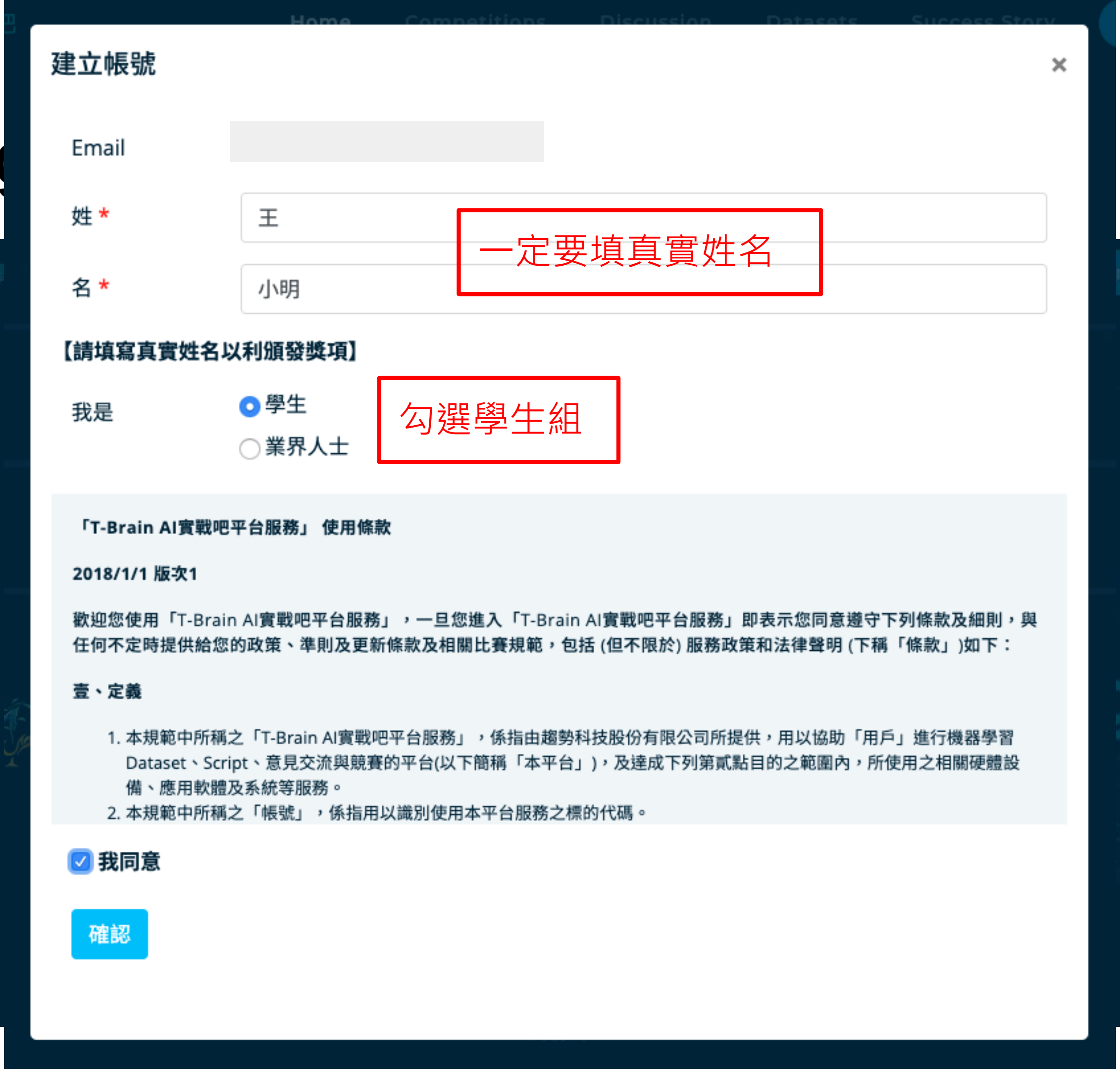
- If you have any problems, please check or post at the competition discussion page first !!!
- If you have any problems, feel free to contact TAs.

- TA in charge: 張緣彩
  - TA hour: Fri. 11:00 ~ 12:00
  - Email: r07922141@ntu.edu.tw

How to register at TBrain platform?



How to reg



建立帳號

×

Email

姓 * 王

名 * 小明

一定要填真實姓名

【請填寫真實姓名以利頒發獎項】

我是 ● 學生 ○ 業界人士

勾選學生組

「T-Brain AI實戰吧平台服務」 使用條款

2018/1/1 版次1

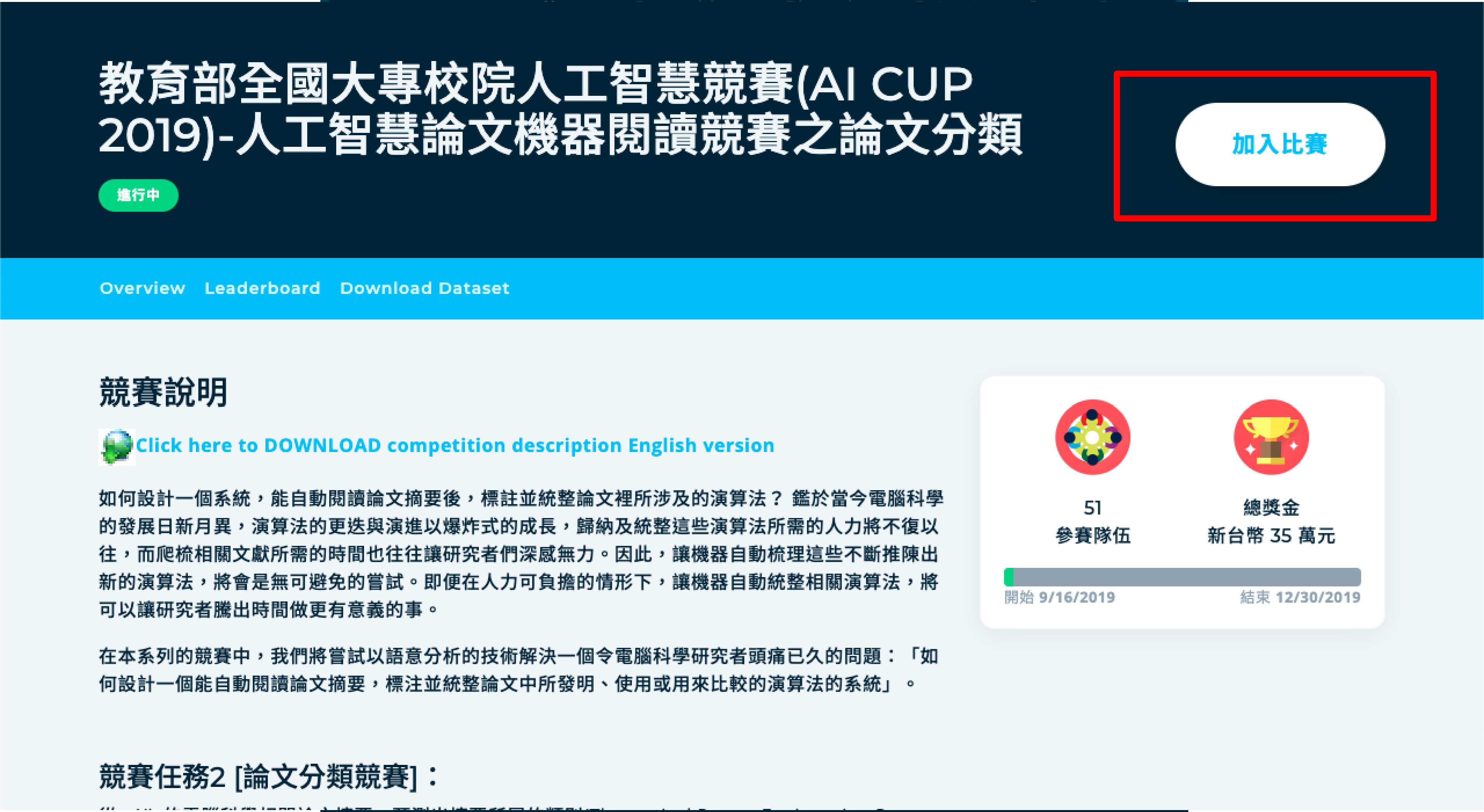
歡迎您使用「T-Brain AI實戰吧平台服務」，一旦您進入「T-Brain AI實戰吧平台服務」即表示您同意遵守下列條款及細則，與任何不定時提供給您的政策、準則及更新條款及相關比賽規範，包括 (但不限於) 服務政策和法律聲明 (下稱「條款」)如下：

壹、定義

1. 本規範中所稱之「T-Brain AI實戰吧平台服務」，係指由趨勢科技股份有限公司所提供，用以協助「用戶」進行機器學習Dataset、Script、意見交流與競賽的平台(以下簡稱「本平台」)，及達成下列第貳點目的之範圍內，所使用之相關硬體設備、應用軟體及系統等服務。
2. 本規範中所稱之「帳號」，係指用以識別使用本平台服務之標的代碼。

☑ 我同意

確認



教育部全國大專校院人工智慧競賽(AI CUP 2019)-人工智慧論文機器閱讀競賽之論文分類

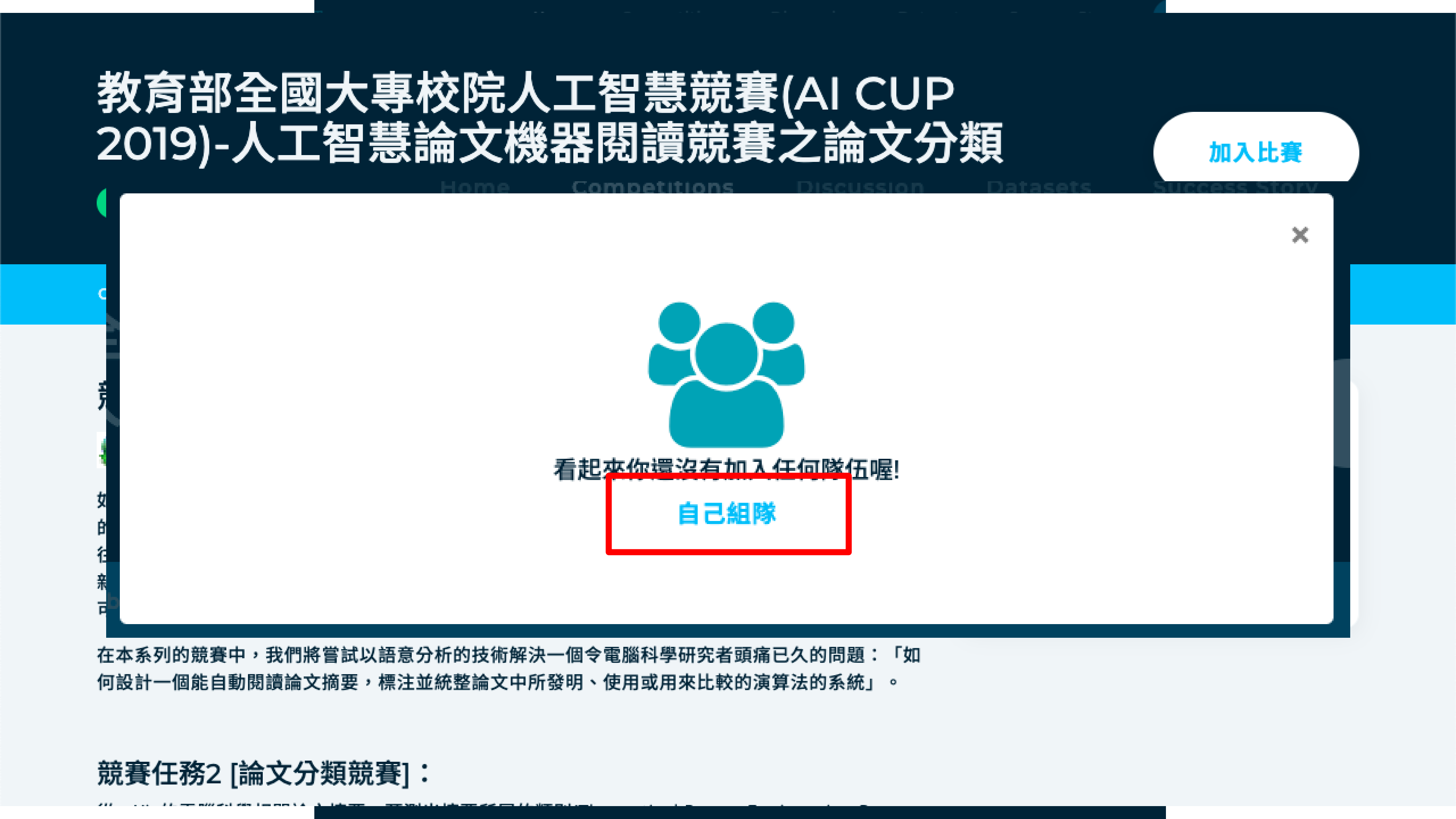
競賽說明

Click here to DOWNLOAD competition description English version

如何設計一個系統，能自動閱讀論文摘要後，標註並統整論文裡所涉及的演算法？ 鑑於當今電腦科學的發展日新月異，演算法的更迭與演進以爆炸式的成長，歸納及統整這些演算法所需的人力將不復以往，而爬梳相關文獻所需的時間也往往讓研究者們深感無力。因此，讓機器自動梳理這些不斷推陳出新的演算法，將會是無可避免的嘗試。即便在人力可負擔的情形下，讓機器自動統整相關演算法，將可以讓研究者騰出時間做更有意義的事。

在本系列的競賽中，我們將嘗試以語意分析的技術解決一個令電腦科學研究者頭痛已久的問題：「如何設計一個能自動閱讀論文摘要，標注並統整論文中所發明、使用或用來比較的演算法的系統」 。

競賽任務2 [論文分類競賽]：

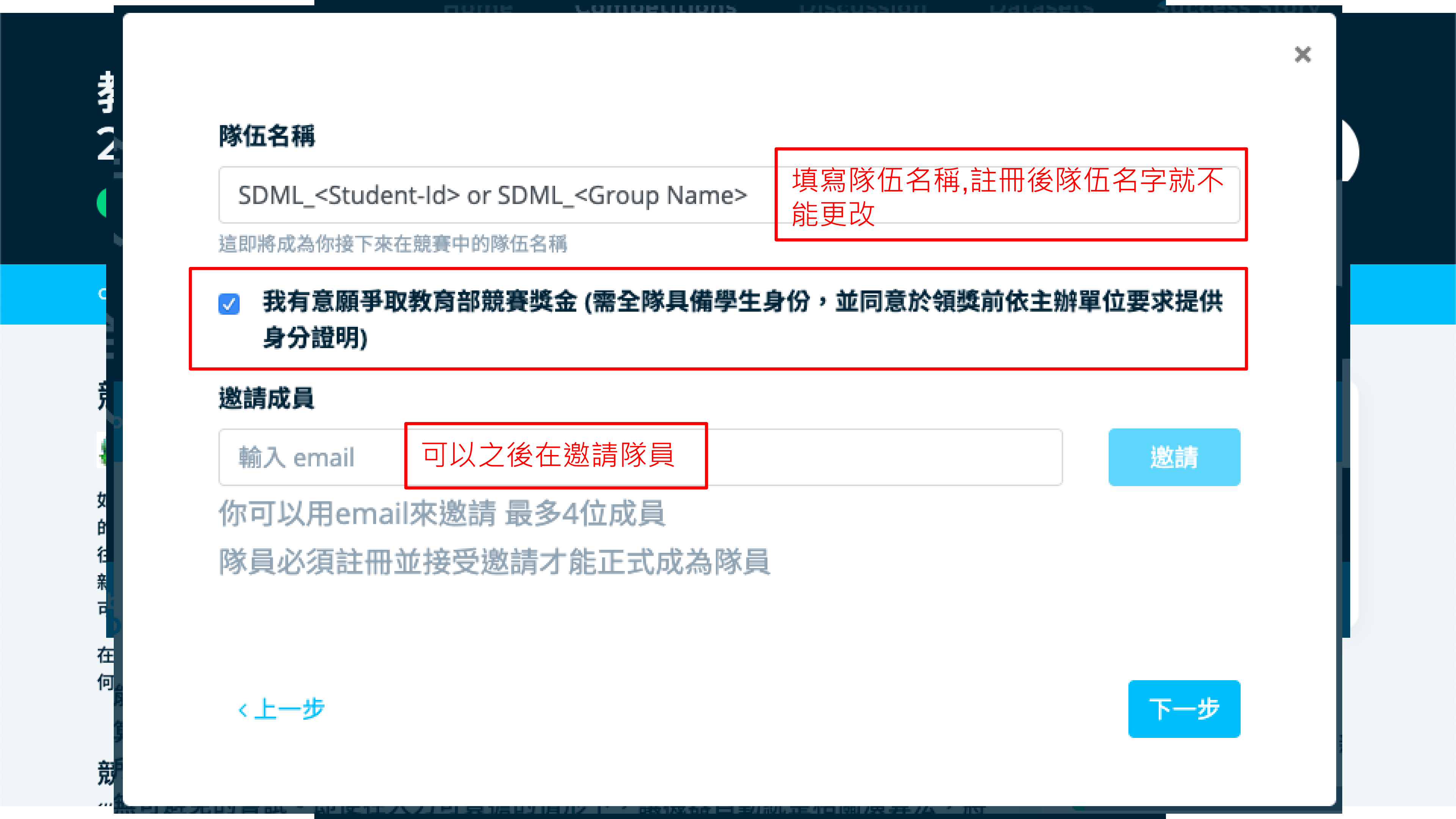


教育部全國大專校院人工智慧競賽(AI CUP 2019)-人工智慧論文機器閱讀競賽之論文分類

看起來你還沒有加入任何隊伍喔!

在本系列的競賽中，我們將嘗試以語意分析的技術解決一個令電腦科學研究者頭痛已久的問題：「如何設計一個能自動閱讀論文摘要，標注並統整論文中所發明、使用或用來比較的演算法的系統」。

競賽任務2 [論文分類競賽]：



隊伍名稱

SDML_<Student-Id> or SDML_<Group Name>

這即將成為你接下來在競賽中的隊伍名稱

填寫隊伍名稱,註冊後隊伍名字就不能更改

☑ 我有意願爭取教育部競賽獎金 (需全隊具備學生身份，並同意於領獎前依主辦單位要求提供身分證明)

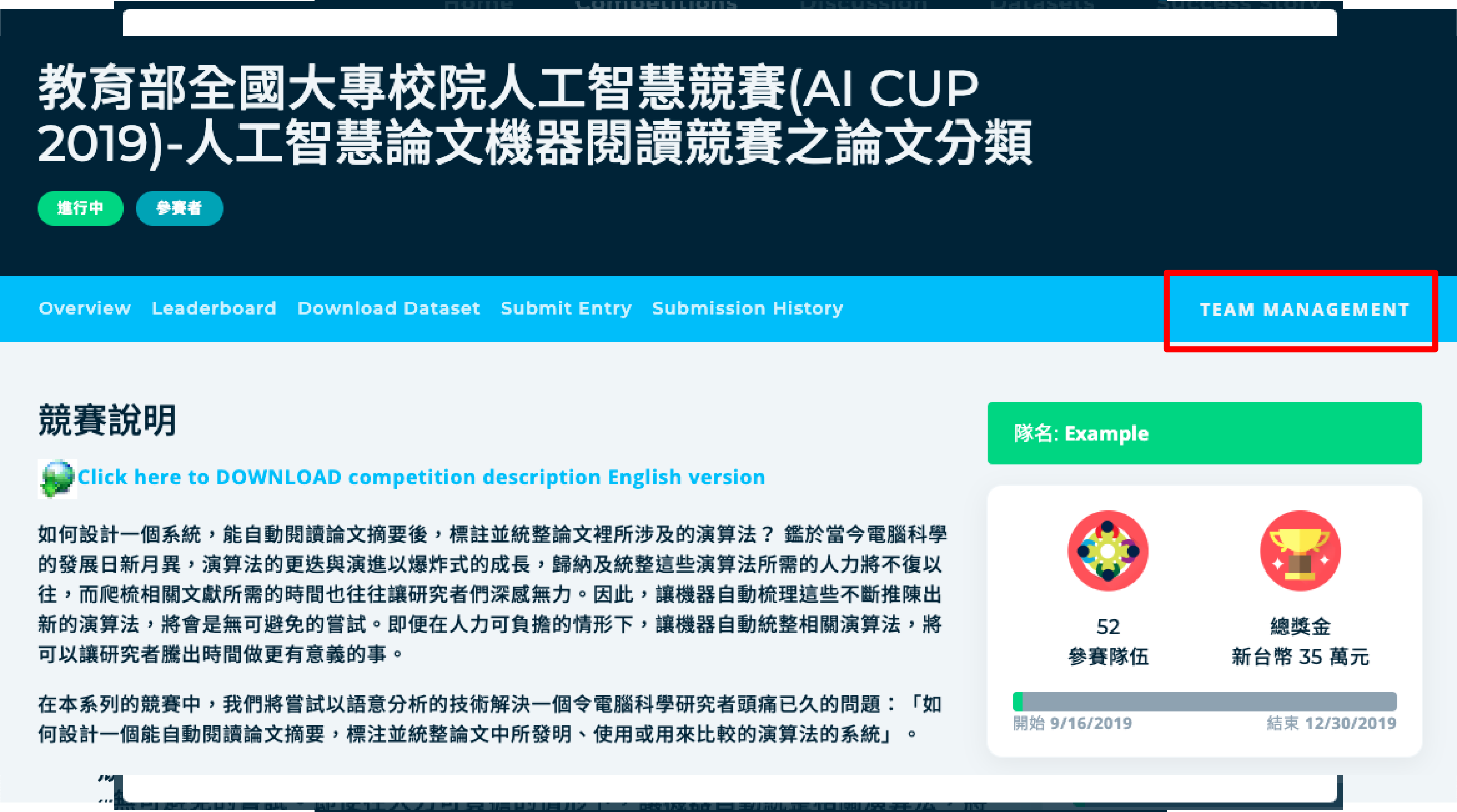
邀請成員

輸入 email

可以之後在邀請隊員

你可以用email來邀請 最多4位成員

隊員必須註冊並接受邀請才能正式成為隊員



教育部全國大專校院人工智慧競賽(AI CUP 2019)-人工智慧論文機器閱讀競賽之論文分類

競賽說明

Click here to DOWNLOAD competition description English version

如何設計一個系統，能自動閱讀論文摘要後，標註並統整論文裡所涉及的演算法？ 鑑於當今電腦科學的發展日新月異，演算法的更迭與演進以爆炸式的成長，歸納及統整這些演算法所需的人力將不復以往，而爬梳相關文獻所需的時間也往往讓研究者們深感無力。因此，讓機器自動梳理這些不斷推陳出新的演算法，將會是無可避免的嘗試。即便在人力可負擔的情形下，讓機器自動統整相關演算法，將可以讓研究者騰出時間做更有意義的事。

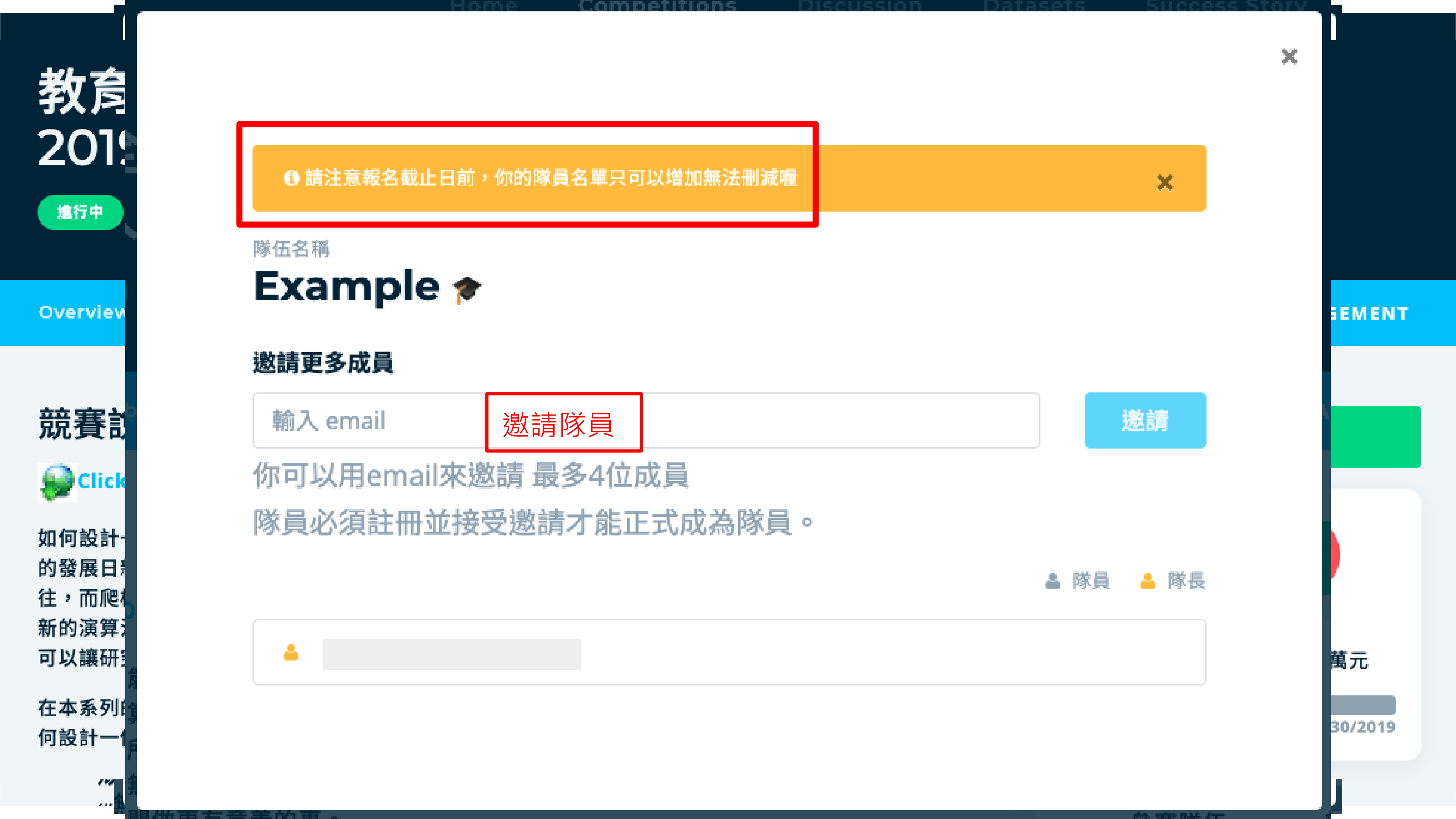
在本系列的競賽中，我們將嘗試以語意分析的技術解決一個令電腦科學研究者頭痛已久的問題：「如何設計一個能自動閱讀論文摘要，標註並統整論文中所發明、使用或用來比較的演算法的系統」。

隊名: Example

52
參賽隊伍

總獎金
新台幣 35 萬元

開始 9/16/2019
結束 12/30/2019



ℹ 請注意報名截止日前，你的隊員名單只可以增加無法刪減喔 ×

隊伍名稱

Example 🎓

邀請更多成員

輸入 email

邀請隊員

邀請

你可以用email來邀請 最多4位成員

隊員必須註冊並接受邀請才能正式成為隊員。

👤 隊員 👤 隊長