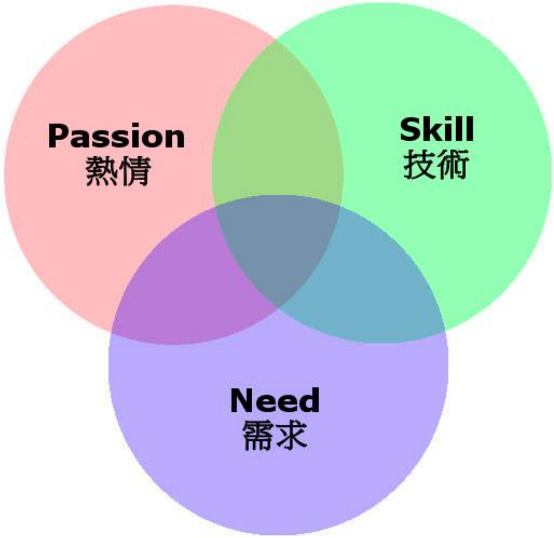
Data Analytics Research Team (DART) – 2025

陳弘軒 Hung-Hsuan Chen
Computer Science & Information Engineering
National Central University

Mission: data science for the society

- Discover the necessity and problem (**Need**)
- Equip with programing and math skills along with domain knowledge to solve the problem (**skill**)
- Willing to practice and make it happen (**Passion**)
- 研究應有所本，不單為研究而研究
- 及早開始研究對學生的好處：產生學生時代的「代表作」
  - 好的論文有助於申請出國留學、好的專案有助於求職



Recent research/project direction

- Develop machine learning models that are
  - Faster (shorter training or inference time)
  - More accurate
  - Better (under certain conditions)
- Apply machine learning to applications
  - Smart sport (精準運動科學)
  - Privacy-preserving machine learning
  - Search engines & recommender systems
  - PM2.5 prediction & sensor malfunction prediction
  - Traffic prediction
  - Personality traits and personality prediction
  - Clip search within videos
  - Log analysis

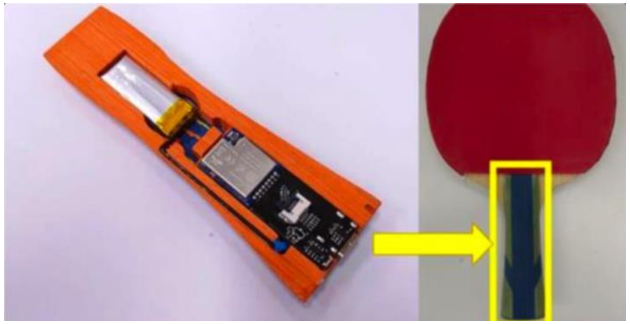
Table of contents

Recent graduate projects

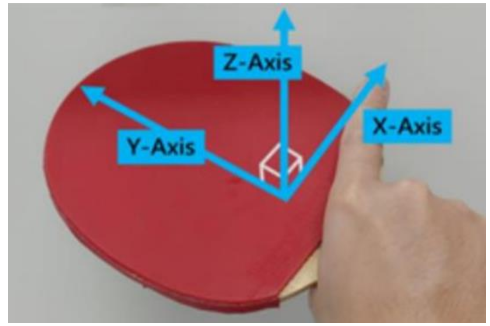
精準運動科學 – 桌球 (2023 – 2027)

- 結合AI的揮拍者能力分析系統
  - 資料蒐集
  - 揮拍者個人運動能力分析
- 戰術分析
  - 資料蒐集與標註
  - 選手打法分析

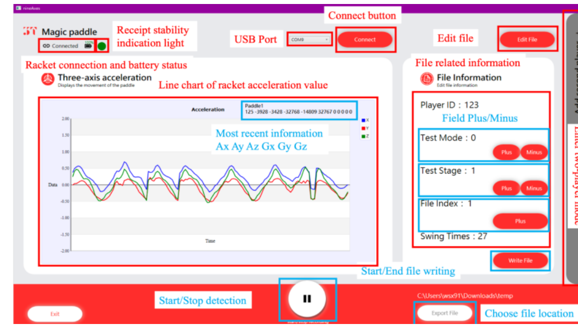
精準運動科學 - 桌球 (結合AI的揮拍者能力分析系統 – 資料蒐集)



智慧球拍內嵌感測器示意圖



x,y,z軸向定義



程式紀錄受測者之揮拍資訊及揮拍模式

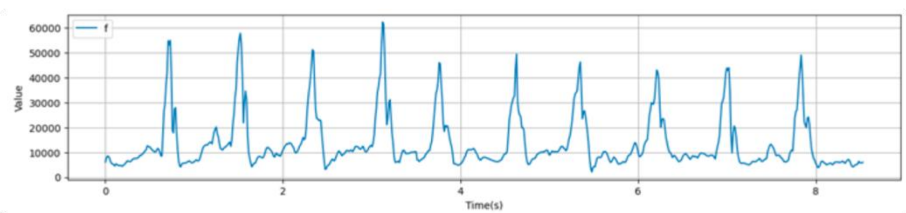
- 單點回擊

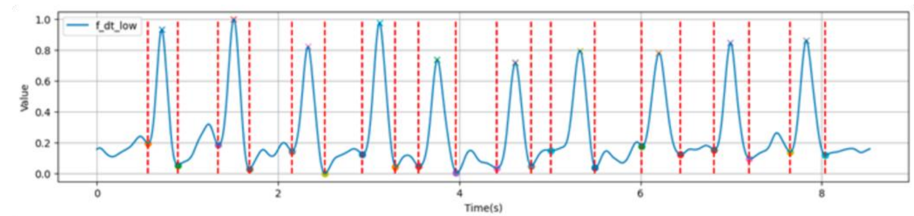
- 兩側移動回擊

C.-Y. Chou, Z.-H. Chen, Y.-H. Sheu, H.-H. Chen, M.-T. Sun and S. K. Wu, **Springer Nature Scientific Data**, 2025

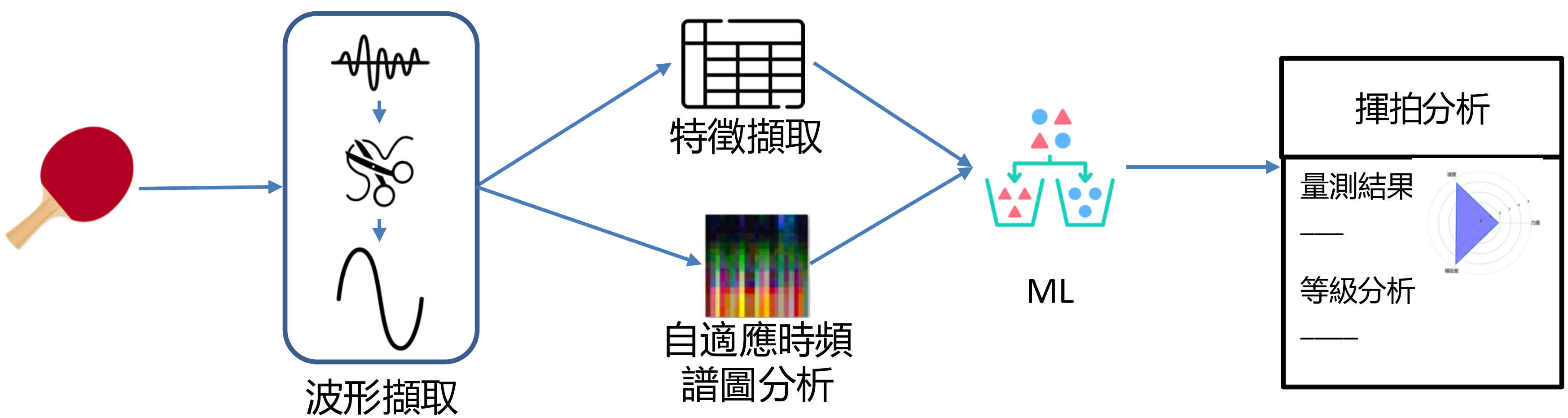
精準運動科學 - 桌球 (結合AI的揮拍者能力分析系統)

- 揮拍波形平滑化技術及自動波形切割

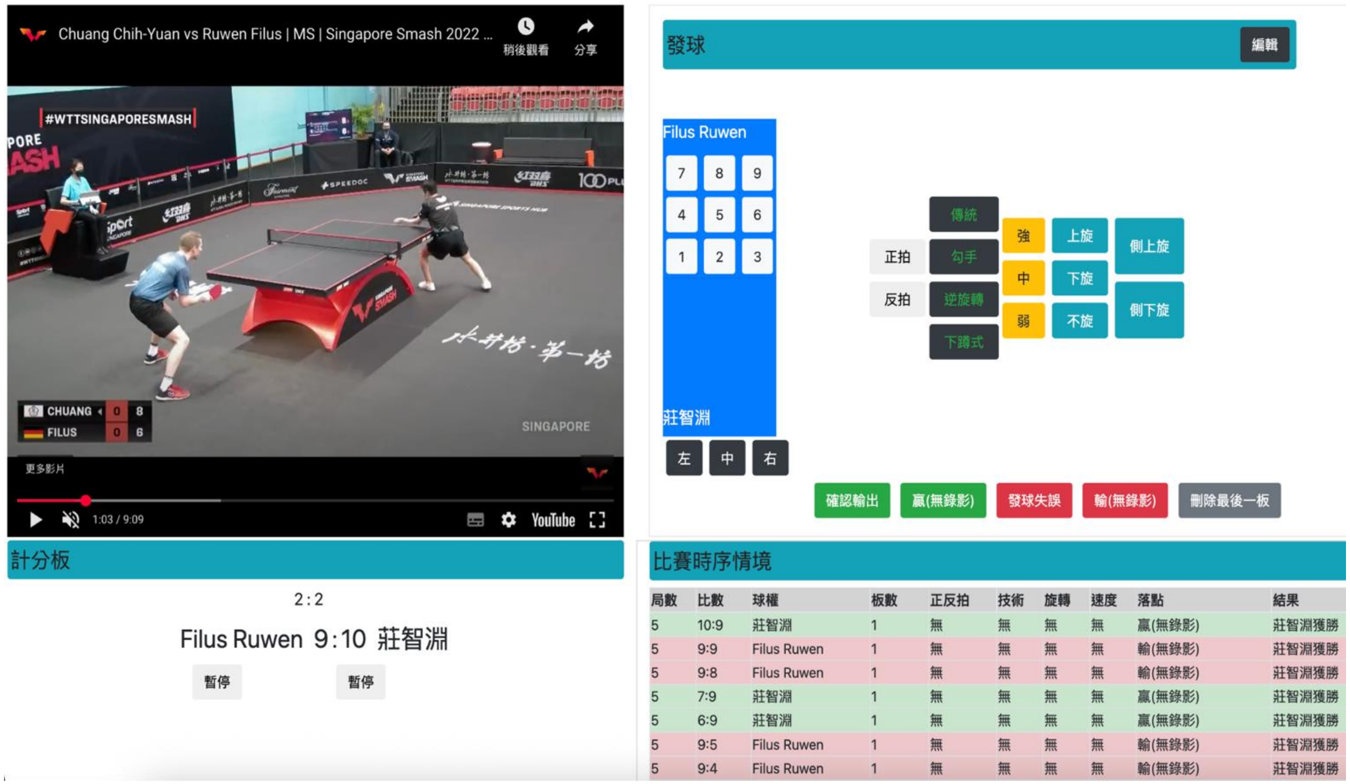




- 揮拍波形AI及大數據分析，產出科學化個人報告：
  (1) 量化揮拍者的揮拍速度、加速度、穩定度數值
  (2) 比較揮拍者的揮拍速度、加速度、穩定度數值在所有受測球員中之百分比
  (3) 揮拍者之等級分析



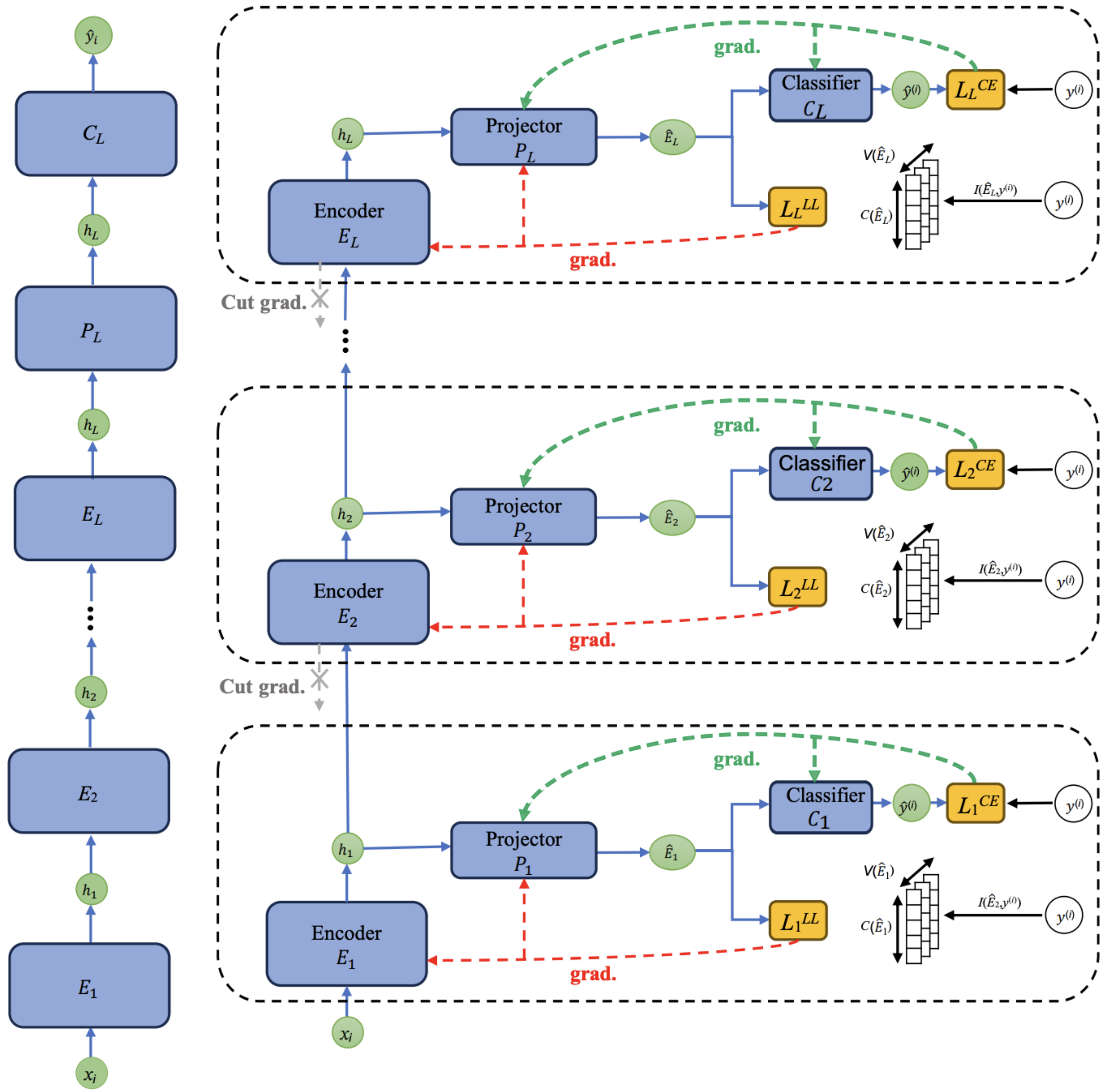
精準運動科學 - 桌球 (戰術分析標註)



- 標註廣泛而詳盡
  - 已標註 100,000+ 拍
  - 涵蓋15個級別賽事 (奧運、世錦賽、亞運、世大運、全國賽、大專盃等)
  - 選手涵蓋29個國家
  - 每拍包括：擊球種類 (如：正反手，發球、攻球、擰球)、落點等，並有原始影片

DeInfoReg – 實現 backprop 的分段平行化 (2023 – 2025)

- Realizes model parallelism for deep learning; maintains high test accuracies across different networks and open datasets



Z.-H. Huang, Y.-T. Lin, H.-H. Chen. **Neurocomputing** 2025.

DeInfoReg – accurate, efficient, and robust

Accurate predictions

Batch Size	64	128	256
IMDB			
BP	87.91 ± 0.44	87.54 ± 0.29	87.50 ± 0.42
AL	87.93 ± 0.37	86.89 ± 0.73	87.98 ± 0.48
SCPL	88.16 ± 0.48	88.89 ± 0.29	88.69 ± 0.32
DeInfoReg	**89.02 ± 0.17**	**89.26 ± 0.09**	**88.86 ± 0.30**
AGNews			
BP	91.05 ± 0.10	90.84 ± 0.11	90.74 ± 0.16
AL	88.20 ± 1.60	86.20 ± 1.90	89.41 ± 0.86
SCPL	91.36 ± 0.21	91.64 ± 0.07	91.66 ± 0.21
DeInfoReg	**91.91 ± 0.13**	**91.95 ± 0.06**	**91.90 ± 0.15**
DBpedia			
BP	97.66 ± 0.05	97.63 ± 0.01	97.59 ± 0.02
AL	91.20 ± 2.37	92.81 ± 2.02	93.58 ± 1.62
SCPL	97.35 ± 0.27	97.47 ± 0.04	97.58 ± 0.05
DeInfoReg	**97.84 ± 0.02**	**97.84 ± 0.02**	**97.85 ± 0.03**

Fast training with multiple GPUs

Batch size	256	512	1024
BP	1x (28.74 sec)	1x (28.44 sec)	1x (28.8 sec)
DeInfoReg (1 GPU)	0.81x	0.85x	0.90x
DeInfoReg (2 GPUs)	1.30x	1.35x	1.35x
DeInfoReg (4 GPUs)	1.41x	1.44x	1.47x

Robust to noisy data

θ: noisy ratio

θ	0.0	0.2	0.4	0.6	0.8
BP	88.27 ± 0.48	83.01 ± 1.2	76.08 ± 0.13	67.71 ± 0.4	57.94 ± 1.75
DeInfoReg	90.15 ± 0.18	85.46 ± 0.66	81.28 ± 1.1	76.36 ± 5.18	68.1 ± 8.31

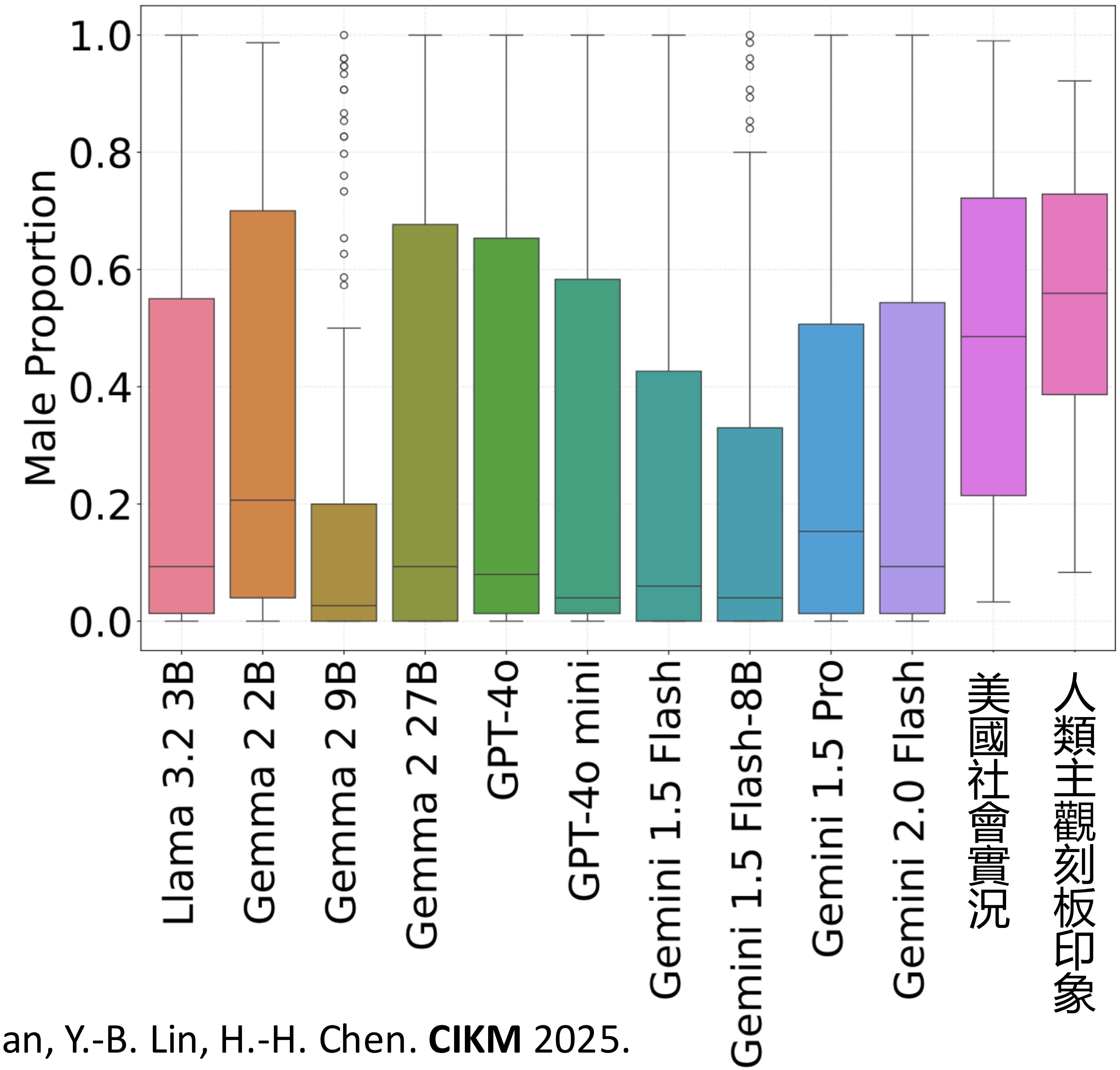
Z.-H. Huang, Y.-T. Lin, H.-H. Chen. **Neurocomputing** 2025.

量化LLM對「職業-性別」的刻板印象 (2024 – 2025)

- 大規模量化LLM對職業的性別刻板印象
  - 10 LLMs (include proprietary & open source)
  - 106個職業
- 發現1: LLM 嚴重高估各職業的女性比例
- 發現2: 即便如此，人類對於「職業-性別」的刻板印象仍存在於 LLM 中
- 發現3: LLM高估女性的原因可能來自 SFT 與 RLHF

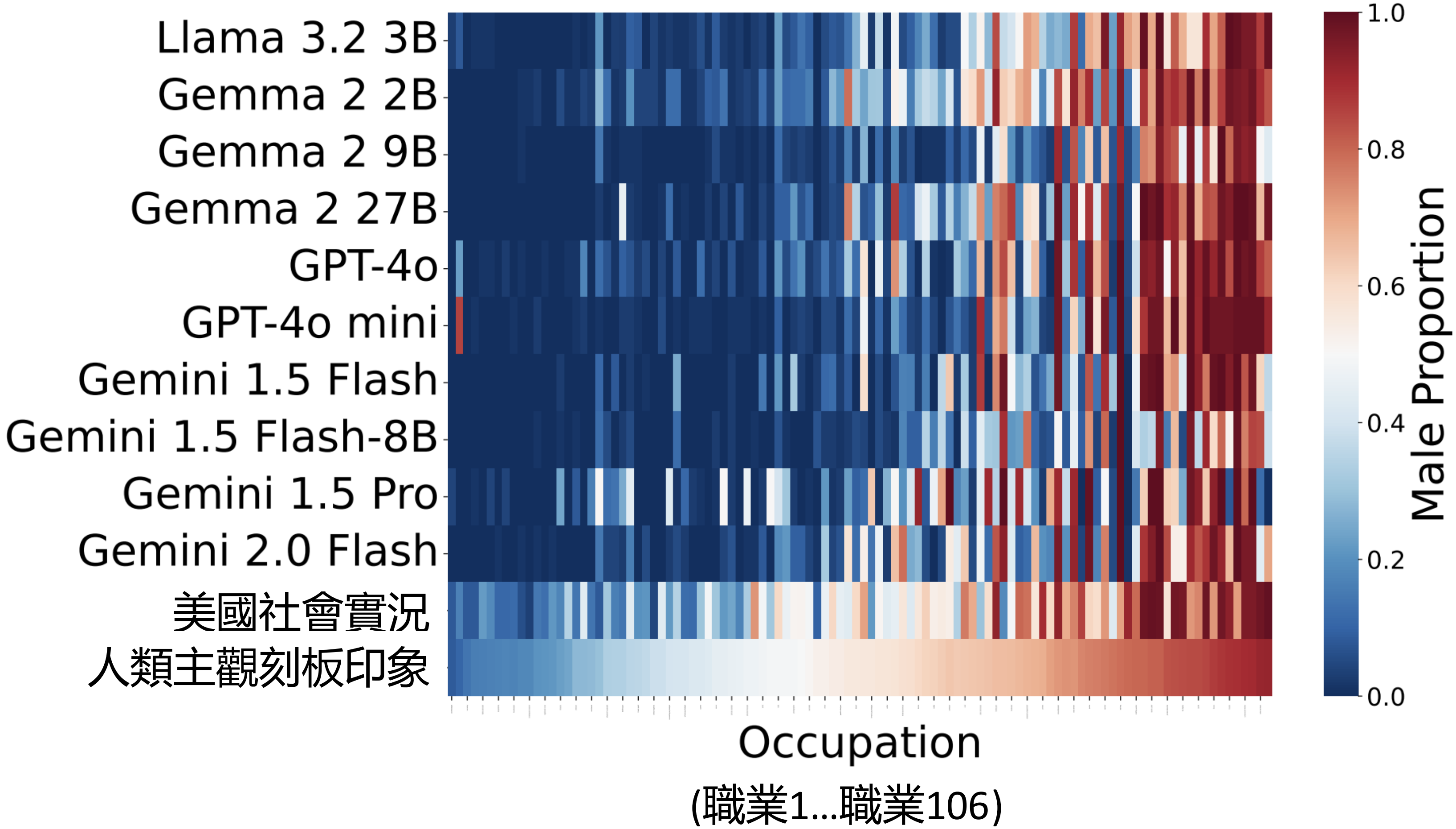
E. Chen, R.-J. Zhan, Y.-B. Lin, H.-H. Chen. **CIKM** 2025.

LLM對「職業-性別」的刻板印象 – 高估各職業的女性比例 (1/2)



E. Chen, R.-J. Zhan, Y.-B. Lin, H.-H. Chen. **CIKM** 2025.

LLM對「職業-性別」的刻板印象 – 高估各職業的女性比例 (2/2)



E. Chen, R.-J. Zhan, Y.-B. Lin, H.-H. Chen. **CIKM** 2025.

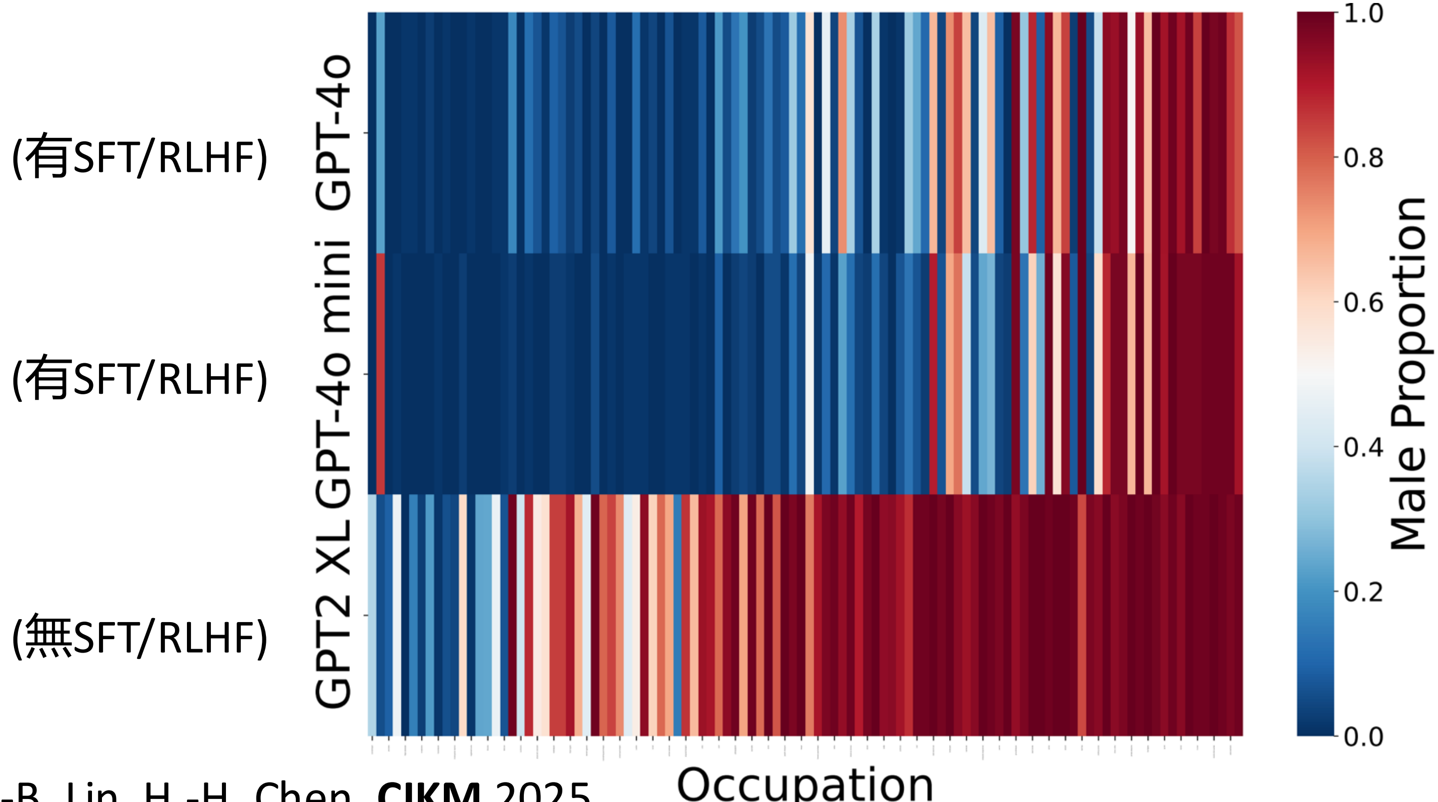
以Gender Ratio 排序職業時，LLM與人類刻板印象高度一致

Model	Kendall's Tau	
	美國社會實況	人類主觀刻板印象
Llama3.2 3B	0.6352	**0.6975**
Gemma2 2B	0.6642	**0.7411**
Gemma2 9B	0.6125	**0.6919**
Gemma2 27B	0.6159	**0.7201**
Gpt-4o mini	0.5939	**0.6712**
Gpt-4o	0.5375	**0.5666**
Gemini 1.5 flash	0.5522	**0.6364**
Gemini 1.5 flash-8B	0.5577	**0.6368**
Gemini 1.5 pro	0.3368	**0.4557**
Gemini 2.0 flash	0.5383	**0.6362**

E. Chen, R.-J. Zhan, Y.-B. Lin, H.-H. Chen. **CIKM** 2025.

LLM高估各職業的女性比例，但職業按 gender ratio 排序卻與人類刻板印象一致?

- 可能成因：Supervised finetuning (SFT) 與 Reinforcement Learning from Human Feedback (RLHF)



E. Chen, R.-J. Zhan, Y.-B. Lin, H.-H. Chen. **CIKM** 2025.

隱私強化技術：差分隱私與合成資料 (2024)

- 技術檢測項目
- 隱私強化技術演算法適用性分析
- 概念驗證實作

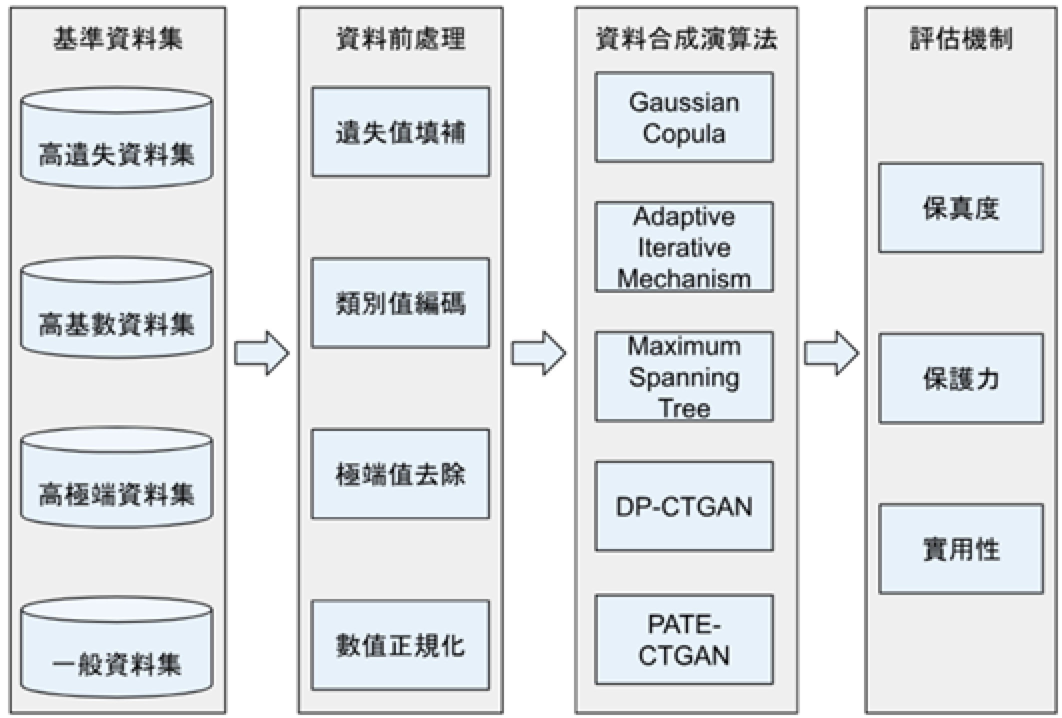
隱私強化技術演算法適用性分析 – 基準資料集特性

高遺失:

- 資料內總遺失筆數超過總筆數15%，或
- 內部遺失值超過 15% 的欄位 (column) 超過總欄位數之 1/3

高基數：類別變項基數 (cardinality) 大於 10 的欄位超過總欄位數之 1/3

高極端：偏態係數 (skewness) 絕對值大於 3 的欄位超過總欄位數之 1/3



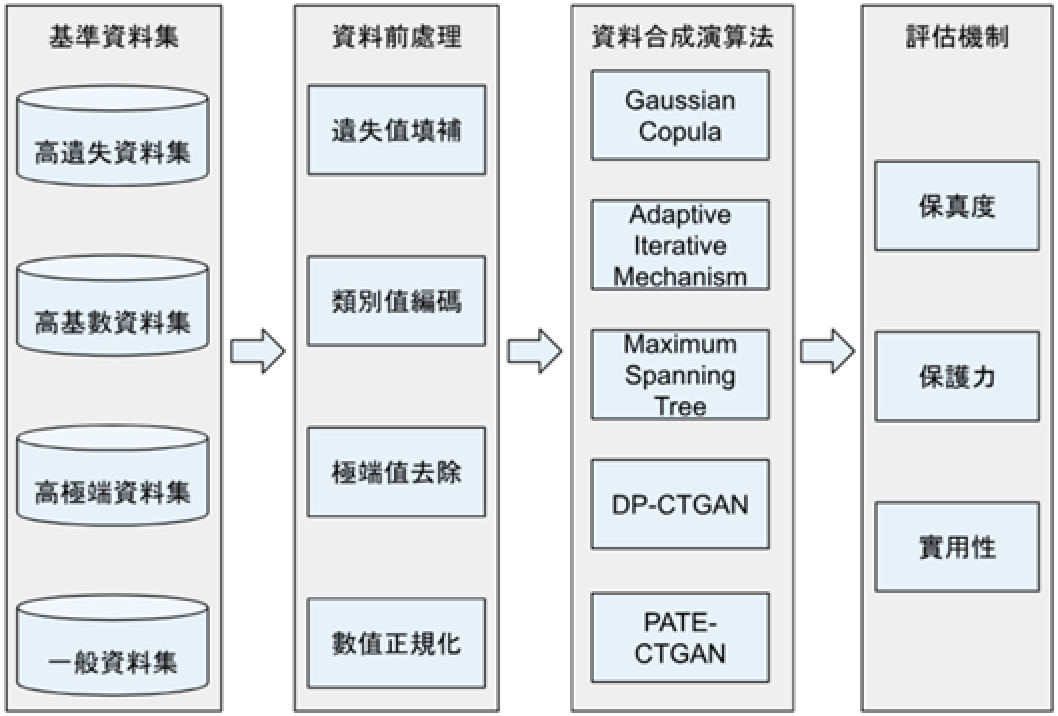
隱私強化技術演算法適用性分析－資料前處理

遺失值填補：填補mean/median/mode或移除

類別值編碼：獨熱/均勻/標籤編碼

極端值去除：由 z-score/四分位距/隔離森林/局部離群因子來決定極端值

數值標準化：scale to [0,1], scale to [-1,1], standardize



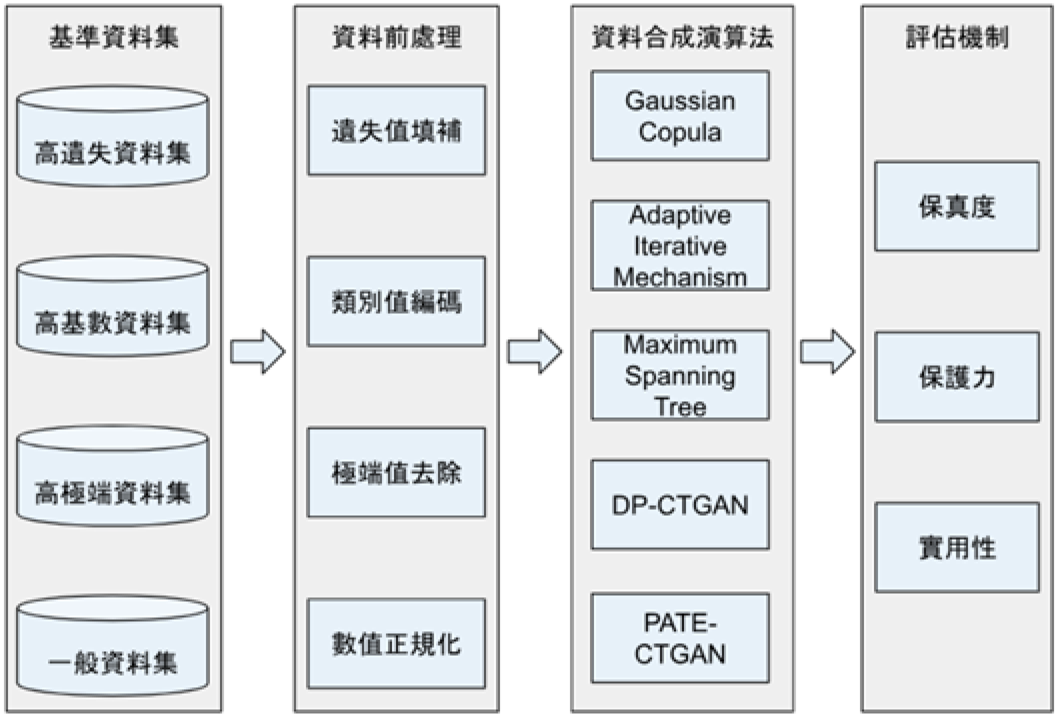
隱私強化技術演算法適用性分析－資料合成演算法

有 DP 保護

- Adaptive Iterative Mechanism (AIM), Maximum Spanning Tree (MST), DP-CTGAN, PATE-CTGAN

無 DP 保護

- Gaussian Copula



隱私強化技術演算法適用性分析 – 評估機制

保真度：合成資料是否能夠真實反映原始資料的特性

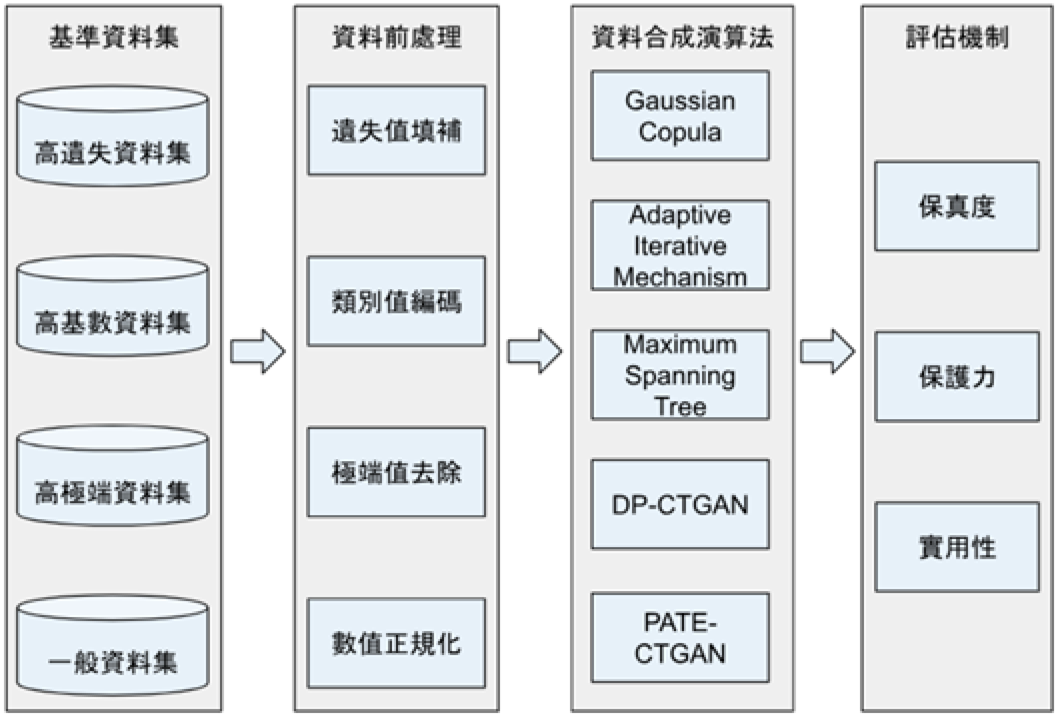
- 單欄統計特性、欄位關聯性等

保護力：合成資料承受攻擊的能力

- 指認性、連結性、推論性風險

實用性：合成資料與原始資料面對下游機器學習任務的成效差距

- 分類：Accuracy, AUROC, F1
- 迴歸：RMSE、R2
- 聚類：輪廓係數



隱私強化技術演算法適用性分析－前處理指引表

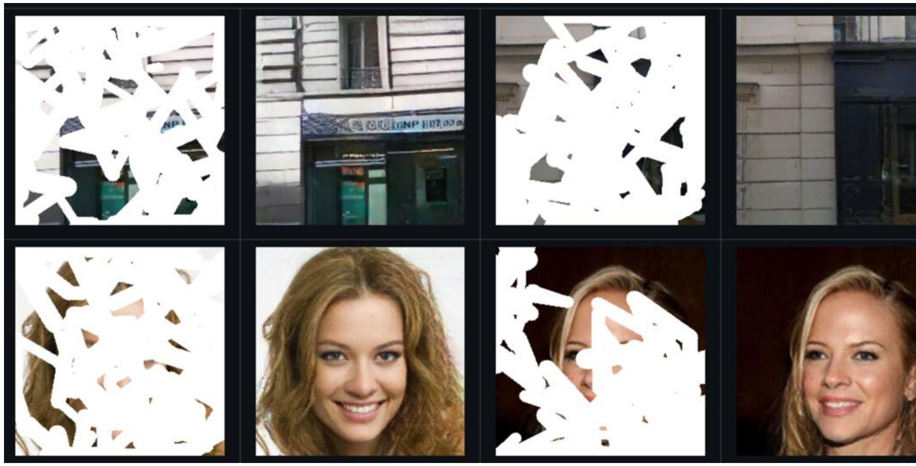
	重視保護力	重視保真度	重視實用性
面對高遺失資料集 (遺失值為數值)	使用 missing_median 處理遺失值	使用 missing_median 處理遺失值	使用 missing_median 處理遺失值
面對高遺失資料集 (遺失值為類別值)	使用 missing_mode 處理遺失值	使用 missing_drop 處理遺失值	使用 missing_mode 處理遺失值
面對高基數資料集	使用 encoder_label 對類別值編碼	使用 encoder_uniform 對類別值編碼	使用 encoder_label 對類別值編碼
面對高極端資料集	無明顯差異， 使用預設方法即可	使用 outlier_isolationforest 或 outlier_lof 處理極端值	使用 outlier_isolationforest 或 outlier_lof 處理極端值
面對一般資料集	無明顯差異， 使用預設方法即可	使用 scaler_minmax 對數值做標準化	無明顯差異， 使用預設方法即可

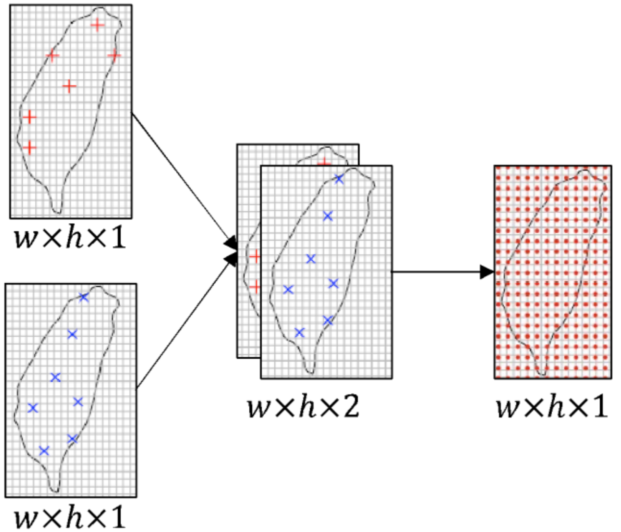
空汙感測網路異質資料整合 (2021 – 2023)

- 合作單位：中大大氣系、太空遙測中心、資工系
- 資工系負責項目：
  - 結合衛星資料與地表空汙感測器共同預測各地空汙

利用圖像修復技術進行多來源PM2.5的融合與補值

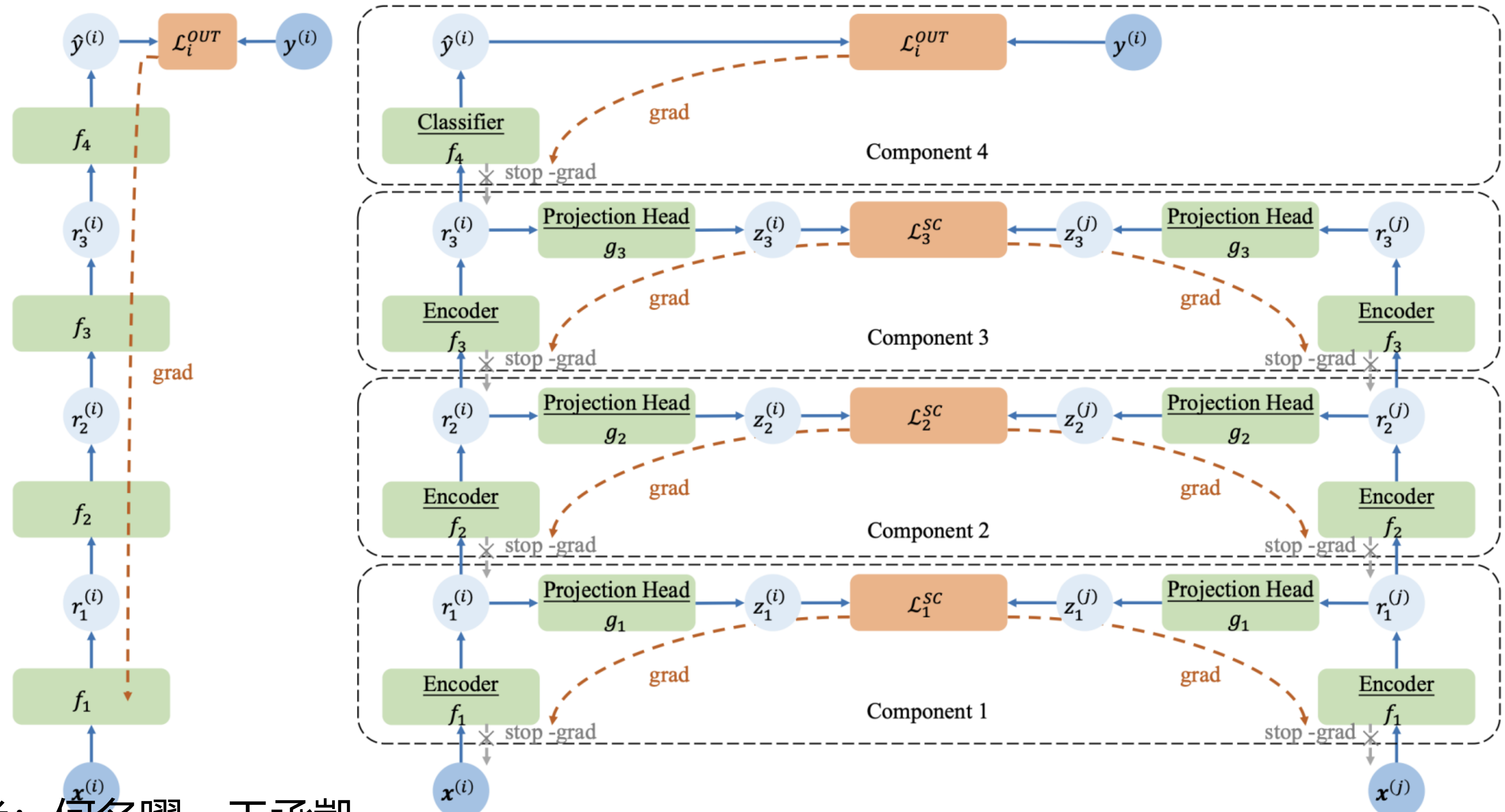
- 圖像修補技術
  - 從部份圖像生成整張圖像
  - 一張圖片有RGB三個頻道
- 多來源PM2.5融合
  - 從部份PM2.5測量生成全台PM2.5濃度地圖
  - 「EPA測站」、「Airbox測站」、及「衛星PM2.5」視為不同頻道
- 相同處
  - 均為三維陣列 (長 x 寬 x 頻道數)
  - 均需無中生有
- 相異處
  - 圖像修補任務中，相同(x,y)坐標點的不同頻道「同時有值」或「同時缺值」
  - PM2.5補值任務中，相同(x,y)坐標點的不同頻道不一定「同時有值」或「同時缺值」





Supervised Contrastive Parallel Learning (SCPL) (1/3)

- Realizes model parallelism for deep learning; maintains high test accuracies across different networks and open datasets



主要作者：何名曜、王承凱

Supervised Contrastive Parallel Learning (SCPL) (2/3)

Standard BP

Device No.	Stage																
GPU0	FW1	FW2	FW3	FW4	LOSS	BW4	BW3		BW2				BW1				UP
Time point	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}	t_{11}	t_{12}	t_{13}	t_{14}	t_{15}	t_{16}	t_{17}

NMP

Device No.	Stage																
GPU0	FW1												BW1				UP
GPU1		FW2							BW2								UP
GPU2			FW3				BW3										UP
GPU3				FW4	LOSS	BW4											UP
Time point	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}	t_{11}	t_{12}	t_{13}	t_{14}	t_{15}	t_{16}	t_{17}

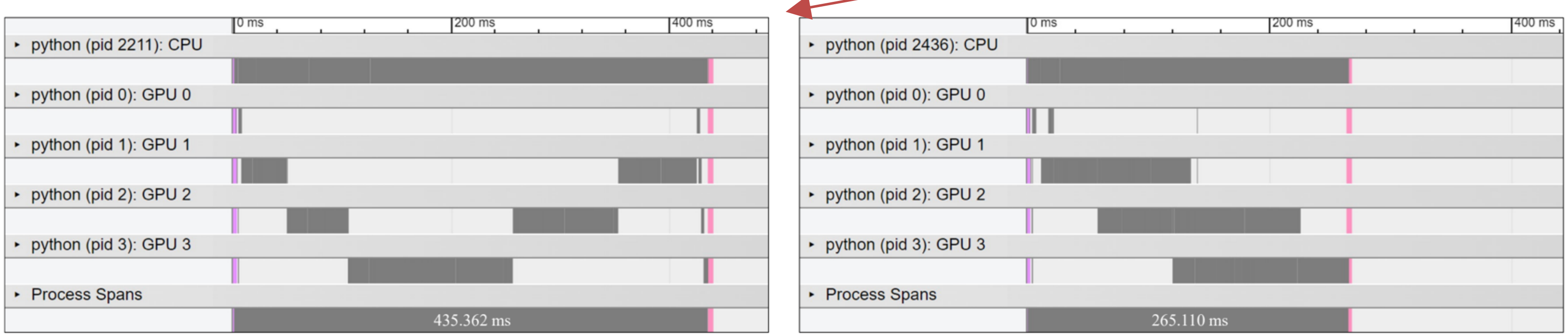
Concept illustration

SCPL

Device No.	Stage							
GPU0	FW1	LOSS	BW1					UP
GPU1		FW2	LOSS	BW2				UP
GPU2			FW3	LOSS	BW3			UP
GPU3				FW4	LOSS	BW4		UP
Time point	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8

FWi: forward for layer i
LOSS: compute loss
BWi : backward for layer i
UP: update parameter values

True training process on 4 GPUs



(a) Training LSTM on IMDB (using NMP). (b) Training LSTM on IMDB (using SCPL).

Supervised Contrastive Parallel Learning (SCPL) (3/3)

Training time speedup ratios (IMDB, transformer)

Batch size	32	64	128	256	512
BP	1x (196 min)	1x (173 min)	1x (156 min)	1x (149 min)	1x (147 min)
GPipe (1 GPU)	0.75x	0.72x	0.72x	0.71x	0.70x
GPipe (2 GPUs)	1.00x	0.92x	0.93x	0.93x	0.92x
GPipe (4 GPUs)	1.35x	1.25x	1.17x	1.16x	1.11x
SCPL (1 GPU)	1.12x	1.07x	1.03x	1.03x	1.05x
SCPL (2 GPUs)	1.43x	1.37x	1.32x	1.37x	1.38x
SCPL (4 GPUs)	1.92x	1.82x	1.66x	1.67x	1.66x

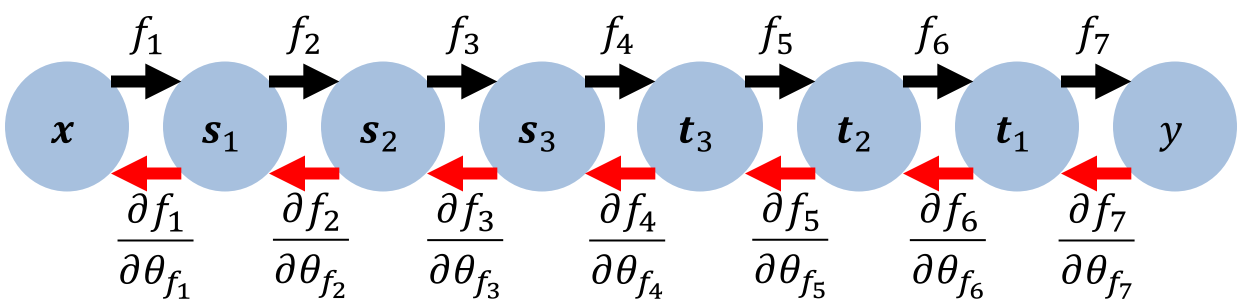
Test accuracies (IMDB)

	LSTM	Transformer
BP	89.68 ± 0.20	87.54 ± 0.44
Early Exit	84.34 ± 0.31	80.24 ± 0.24
AL	86.41 ± 0.61	85.65 ± 0.77
SCPL	$\mathbf{89.84} \pm 0.10$ †	$\mathbf{89.03} \pm 0.12$ †

主要作者：何名曜、王承凱

Associated learning (AL) (1/2)

- AL: an alternative to end-to-end backpropagation
- AL decomposes a network into small components:
  - Each component has a local objective function
  - Parameters in different components can be updated simultaneously
    - Eliminate backward lock, so pipelined training is possible; increase throughput



D. Wu, D.-N. Lin, V. Chen, H.-H. Chen. **ICLR** 2022.
Y.-W. Kao and H.-H. Chen. **MIT Neural Computation** 2021.

Associated learning (2/2)

- Results on image classification (CIFAR-100)

	BP	AL
Vanilla CNN	$26.5 \pm 0.4\%$	$\mathbf{29.7 \pm 0.2\%}$
VGG	$65.8 \pm 0.3\%$	$\mathbf{67.1 \pm 0.3\%}$

- Results on NLP-1 (IMDB)

	BP	AL
LSTM	$88.10 \pm 0.50\%$	$\mathbf{89.04 \pm 0.37\%}$

- Results on NLP-2 (AGNews)

	BP	AL
LSTM	$88.56 \pm 0.97\%$	$\mathbf{91.42 \pm 0.42\%}$

D. Wu, D.-N. Lin, V. Chen, H.-H. Chen. **ICLR** 2022.
Y.-W. Kao and H.-H. Chen. **MIT Neural Computation** 2021.

Parametric spectral clustering (1/2)

- Spectral Clustering
  - Pro: effective for non-linearly clustering
  - Con: high computational costs and memory usage
- Proposed Parametric Spectral Clustering (PSC)
  - Learns to project data into spectral space
  - Supports incremental clustering by updating clusters without retraining the entire model
  - Efficient for real-time scenarios with dynamic datasets

J.-C. Chen and H.-H. Chen. IEEE International Conference on **Big Data** 2023

Parametric spectral clustering (2/2)

- SC vs. PSC: computation cost and accuracy
  - Faster, memory-friendly, little (or no) sacrifice on accuracy

Method	Execution time (s)	Peak memory usage (MB)	ClusterAcc
SC	2462	5331	0.794 ± 0.04
PSC ($r = 1/6$)	Training: 453 (↓ 82%) Inference: 0.443 ± 0.027	Training: 1032 (↓ 81%) Inference: 92.83 ± 1.965	0.732 ± 0.076
PSC ($r = 2/6$)	Training: 533 (↓ 78%) Inference: 0.419 ± 0.088	Training: 1717 (↓ 68%) Inference: 92.63 ± 2.09	0.739 ± 0.089
PSC ($r = 3/6$)	Training: 706 (↓ 71%) Inference: 0.437 ± 0.048	Training: 2532 (↓ 53%) Inference: 96.21 ± 1.785	0.764 ± 0.041
PSC ($r = 4/6$)	Training: 1029 (↓ 58%) Inference: 0.379 ± 0.06	Training: 3328 (↓ 38%) Inference: 96.54 ± 2.271	0.775 ± 0.046
PSC ($r = 5/6$)	Training: 1472 (↓ 40%) Inference: 0.491 ± 0.036	Training: 4937 (↓ 7%) Inference: 96.5 ± 2.15	0.819 ± 0.039

J.-C. Chen and H.-H. Chen. IEEE International Conference on **Big Data** 2023

偵測低調的網軍 (1/3)

- 電腦容易偵測高調的網軍
  - 常發言、常回文、常推/噓文等
- 偵測低調的網軍相對困難

AUPRC scores of detecting active and low active spammers

	active users	inactive users	diff
XGBoost	0.8892	0.5157	0.3735
LightGBM	0.7421	0.4888	0.2533
Random Forest	0.8317	0.5147	0.3163

- 但你知道大部份的網軍是「低調」的嗎?

Group	Percentile of active value	Active value	# normal accounts	CDF of normal accounts (a)	# spammers	CDF of spammers (b)	(b) − (a)
G_1	[0%, 10%)	0-18	4112	9%	222	24%	15%
G_2	[10%, 20%)	19-45	4418	20%	163	42%	22%
G_3	[20%, 30%)	46-84	4508	30%	86	52%	22%
G_4	[30%, 40%)	85-135	4223	40%	59	58%	18%
G_5	[40%, 50%)	136-211	4453	50%	57	64%	14%
G_6	[50%, 60%)	212-315	4096	59%	76	73%	14%
G_7	[60%, 70%)	316-494	4320	69%	112	85%	16%
G_8	[70%, 80%)	495-817	4368	79%	67	92%	13%
G_9	[80%, 90%)	818-1663	4638	90%	51	98%	8%
G_{10}	[90%, 100%]	$\geq$ 1664	4554	100%	19	100%	0%

R.-Y. Wang and H.-H. Chen. IEEE International Conference on **Web Intelligence** 2023

偵測低調的網軍 (2/3)

- 使用傳統機器學習或深度學習偵測低活躍網軍成效不彰

AUPRC scores of detecting less active and highly active spammers

	[0%, 10%)	[10%, 20%)	[80%, 100%]
XGBoost	0.52 ± 0.01	0.48 ± 0.03	0.89 ± 0.01
LightGBM	0.49 ± 0.02	0.40 ± 0.04	0.74 ± 0.02
Random Forest	0.51 ± 0.03	0.27 ± 0.02	0.83 ± 0.02
Fully Connected	0.35 ± 0.06	0.38 ± 0.05	0.75 ± 0.03
ConvNet	0.17 ± 0.06	0.26 ± 0.14	0.80 ± 0.33
Soft Voting [22]	0.40 ± 0.01	0.43 ± 0.01	0.76 ± 0.01
Hard Voting [22]	0.43 ± 0.02	0.47 ± 0.02	0.70 ± 0.03
Stacking [22]	0.42 ± 0.01	0.47 ± 0.03	0.67 ± 0.01

- GNN模型 vs. 最佳非 GNN 模型： GNN更精確地偵測低活躍網軍

GNN vs. XGBoost (best among non-GNN models)

	[0%, 10%)	[10%, 20%)	[80%, 100%]
XGBoost	0.52 ± 0.01	0.48 ± 0.03	0.89 ± 0.01 †
GCN	$\mathbf{0.66} \pm 0.18$	0.38 ± 0.13	0.72 ± 0.07
TAGCN ($K = 1$)	$\mathbf{0.64} \pm 0.04$	$\mathbf{0.79} \pm 0.06$	$\mathbf{0.89} \pm 0.07$ †
TAGCN ($K = 2$)	$\mathbf{0.68} \pm 0.02$	$\mathbf{0.84} \pm 0.05$ †	$\mathbf{0.89} \pm 0.08$ †
TAGCN ($K = 3$)	$\mathbf{0.71} \pm 0.04$ †	$\mathbf{0.80} \pm 0.07$	$\mathbf{0.89} \pm 0.06$ †
GAT	$\mathbf{0.62} \pm 0.09$	$\mathbf{0.77} \pm 0.05$	$\mathbf{0.89} \pm 0.06$ †

R.-Y. Wang and H.-H. Chen. IEEE International Conference on **Web Intelligence** 2023

偵測低調的網軍 (3/3)

- 加入社群特徵可幫助所有模型更好地偵測網軍

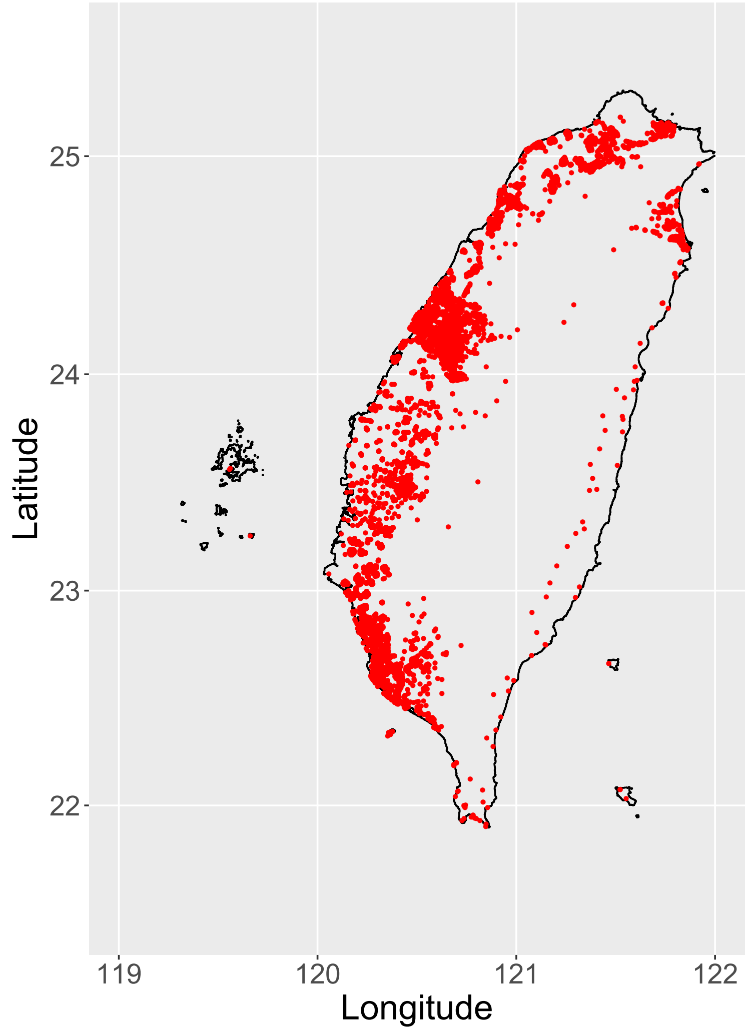
AUPRC scores of the models when including social features

Type	Model	[0%, 10%)	[10%, 20%)	[80%, 100%]	[0%, 100%]
Non-GNN-based models (including social features)	XGBoost	0.83 ± 0.01	$\mathbf{0.74} \pm 0.03$	$\mathbf{0.90} \pm 0.02$	$\mathbf{0.86} \pm 0.00$
	LightGBM	$\mathbf{0.86} \pm 0.02$	0.72 ± 0.05	0.88 ± 0.02	0.82 ± 0.00
	Random Forest	0.85 ± 0.01	0.56 ± 0.05	0.85 ± 0.02	0.79 ± 0.00
	Fully Connected	0.53 ± 0.07	0.51 ± 0.06	0.76 ± 0.05	0.64 ± 0.04
	ConvNet	0.43 ± 0.09	0.68 ± 0.07	0.83 ± 0.04	0.66 ± 0.06
	Soft Voting [22]	0.69 ± 0.00	0.56 ± 0.01	0.76 ± 0.01	0.72 ± 0.00
	Hard Voting [22]	0.67 ± 0.01	0.63 ± 0.02	0.70 ± 0.03	0.74 ± 0.01
	Stacking [22]	0.54 ± 0.02	0.56 ± 0.03	0.67 ± 0.01	0.69 ± 0.02
GNN-based models (including social features)	GCN	0.62 ± 0.08	0.52 ± 0.05	0.83 ± 0.08	0.69 ± 0.03
	TAGCN ($K = 1$)	0.79 ± 0.03	$\mathbf{0.97} \pm 0.05$	$\mathbf{0.99} \pm 0.04$	$\mathbf{0.92} \pm 0.01$
	TAGCN ($K = 2$)	0.82 ± 0.03	$\mathbf{0.98} \pm 0.02$	$\mathbf{0.99} \pm 0.03$	$\mathbf{0.93} \pm 0.02$
	TAGCN ($K = 3$)	0.85 ± 0.02	$\mathbf{0.98} \pm 0.03$	$\mathbf{0.98} \pm 0.01$	$\mathbf{0.94} \pm 0.01$
	GAT	0.73 ± 0.06	$\mathbf{0.91} \pm 0.06$	$\mathbf{0.92} \pm 0.07$	$\mathbf{0.87} \pm 0.05$

R.-Y. Wang and H.-H. Chen. IEEE International Conference on **Web Intelligence** 2023

空汙感測器故障預測 – supervised learning-based

- 10,000+ 空汙感測器 (in 2021)，但有相當比例之量測值不精準
- 採定期巡檢，但人力成本極高
- 智慧巡檢：以圖卷積網路 (Graphical Convolutional Network) 與時間卷積網路整合時空資訊預測故障之感測器
- 訓練資料採用 2018 年的部份資料
- 工研院於2018年5月至12月巡檢144個測站，以巡檢結果做為測試資料
    - 28個異常
    - 116個正常
    - 我們以此巡檢紀錄評估各種異常偵測演算法的優劣



D. Wu, T.-H. Lin, X.-R. Zhang, C.-P. Chen, J.-H. Chen, H.-H. Chen. **IEEE Sensors Journal** 23(15), 2023 (**Featured article**)
T.-H. Lin, X.-R. Zhang, C.-P. Chen, J.-H. Chen, H.-H. Chen. **IEEE Sensors Journal** 22(3), 2022

實驗結果 – AUROC 分數

Type	Model	ROC mean	ROC std
Rule based	ADF-5 (5 是 [6] 中給的超參數值)	0.624	0.0
	ADF-10(ROC Best)	0.694	0.0
ML(無圖卷積)	Random Forest	0.6878	0.006261
	Lasso	0.7000	0.015652
	Ridge	0.7085	0.013472
	TCN	0.7066	0.007701
	DNN	0.6940	0.007211
	LSTM	0.7090	0.007211
ML(圖卷積)	GraphWaveNet	**0.7260**	**0.010826**
	STGCN	0.7214	0.018569

D. Wu, T.-H. Lin, X.-R. Zhang, C.-P. Chen, J.-H. Chen, H.-H. Chen. **IEEE Sensors Journal** 23(15), 2023 (**Featured article**)
T.-H. Lin, X.-R. Zhang, C.-P. Chen, J.-H. Chen, H.-H. Chen. **IEEE Sensors Journal** 22(3), 2022

實驗結果 – Precision@*k*

- Precision@*k*: 若按建議依序檢查*k*個測站，實際有問題的測站在*k*個測站中的佔比

Type	Model	P@10	P@20	P@30	P@40	P@50
隨機巡檢		0.194	0.194	0.194	0.194	0.194
Rule based	ADF-5	0.300	0.350	0.270	0.330	0.320
	ADF-10(ROC Best)	0.500	0.500	0.400	0.380	0.320
ML(無圖卷積)	Random Forest	0.380	0.370	0.400	0.342	0.320
	Lasso	0.580	0.430	0.394	0.370	0.320
	Ridge	0.600	0.433	0.395	0.375	0.337
	TCN	0.600	0.410	0.412	0.338	0.320
	DNN	0.500	0.430	0.374	0.344	0.312
	LSTM	0.600	0.410	0.368	0.332	0.336
ML(圖卷積)	GraphWaveNet	0.600	0.417	**0.417**	0.380	0.353
	STGCN	**0.640**	**0.450**	0.398	**0.386**	**0.360**

D. Wu, T.-H. Lin, X.-R. Zhang, C.-P. Chen, J.-H. Chen, H.-H. Chen. **IEEE Sensors Journal** 23(15), 2023 (**Featured article**)
T.-H. Lin, X.-R. Zhang, C.-P. Chen, J.-H. Chen, H.-H. Chen. **IEEE Sensors Journal** 22(3), 2022

實驗結果 – Recall@*k*

- Recall@*k*: 按建議依序檢查*k*個測站，找出有問題的測站數量與實際有問題測站數量 (28個) 的比值

Type	Model	R@10	R@20	R@30	R@40	R@50
隨機巡檢		0.069	0.139	0.208	0.278	0.347
Rule based	ADF-5	0.110	0.250	0.290	0.460	0.570
	ADF-10(ROC Best)	0.180	0.360	0.430	0.540	0.570
ML(無圖卷積)	Random Forest	0.136	0.266	0.428	0.484	0.570
	Lasso	0.204	0.306	0.422	0.524	0.570
	Ridge	0.210	0.308	0.423	0.533	0.603
	TCN	0.212	0.296	0.442	0.476	0.570
	DNN	0.180	0.308	0.398	0.484	0.560
	LSTM	0.214	0.293	0.394	0.474	0.600
ML(圖卷積)	GraphWaveNet	0.214	0.300	**0.447**	0.543	0.630
	STGCN	**0.230**	**0.322**	0.428	**0.550**	**0.642**

D. Wu, T.-H. Lin, X.-R. Zhang, C.-P. Chen, J.-H. Chen, H.-H. Chen. **IEEE Sensors Journal** 23(15), 2023 (**Featured article**)
T.-H. Lin, X.-R. Zhang, C.-P. Chen, J.-H. Chen, H.-H. Chen. **IEEE Sensors Journal** 22(3), 2022

空汙感測器故障預測 – semi-supervised learning-based

- 10000+ 個空汙感測器中，只有144個有「正常」或「故障」的標準答案
- Fully supervised learning: 僅有 144 筆訓練資料
- Semi-supervised learning：融合有標準答案的資料及其他沒有標準答案的資料共同訓練

主要作者：張欣茹

空汙感測器故障預測 – semi-supervised learning-based

非機器學習模型	隨機巡檢	0.1940 ± 0.0000			
	ADF-5	0.2900 ± 0.0000			
	ADF-10	0.4400 ± 0.0000			
		折線圖	熱力圖	統整性資料	統整及時序資料
監督式模型	linear regression	0.2769	0.3137	0.3339	0.3163
	ridge regression	0.3214	0.3876	0.3337	0.3159
	random forest	0.3290	0.4292	**0.4471**	**0.4588**
	SSDO with iforest	0.3374	0.4555	0.3061	0.2883
	SSDO with COP-kmeans	0.3399	0.5158	0.3177	0.2554
無監督式模型	Isolation fores	0.1886	0.2003	0.2375	0.2578
半監督式模型	SSDO with iforest	0.3712	0.4114	0.2645	0.3773
	SSDO with COP-kmeans	0.3640	0.4162	0.2809	0.3214
	Deep SAD	**0.8099**	**0.8048**	0.3450	0.4215

不同模型在不同資料中所得到的PR-AUC

主要作者：張欣茹

Privacy-preserving machine learning (1/2)

- 政府開放資料可促進創新，但帶有隱私資訊的資料難以直接開放
- 假名化：將可識別個人的資訊 (e.g., 姓名、身份證字號) 替換為假名或代號
- 假名化 (pseudonymization) 仍會洩露隱私
  - 莊智淵是台灣唯一40+的桌球國手
  - 中央大學資工系學生中來自金門高中僅個位數
- 差分隱私 (Differential privacy):
  - (Layman's language) 資料中加入雜訊，使得難以從資料中推測出具體個人的資訊
  - 但差分隱私會讓資料的實用性降低

Privacy-preserving machine learning (2/2)

-

Extended Clickstream

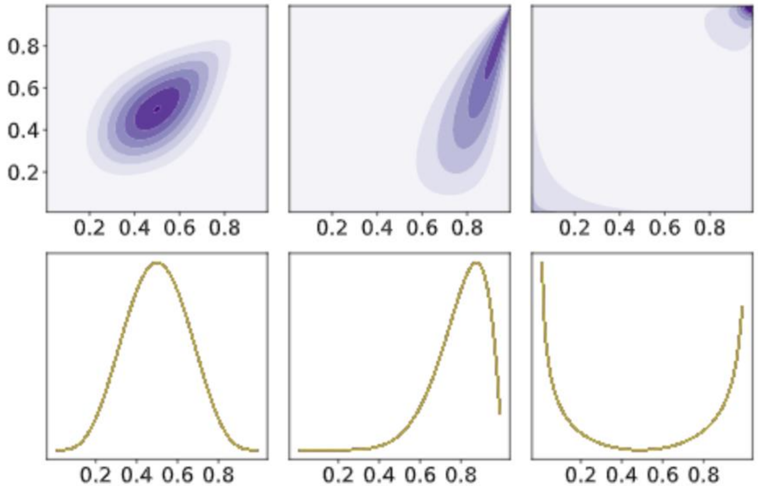
- Weblog approximately records only half of a user's page visits
- 8.1% of the visits recorded in the weblog may not come from a user's conscious actions
- Clickstream is an incomplete collection of users' web visiting

Category	ICS + ECS				CS		ICS		ECS		Rank Diff
	Rank(1)	Count	Perc.(%)	CDF(%)	Rank(2)	Count	Rank	Count	Rank	Count	(1)-(2)
Streaming Media and Download	1	1110256	17.57	17.57	3	558327	2	541878	1	568378	-2
Social Networking	2	929709	14.72	32.29	1	608252	1	591064	2	338645	1
Search Engines and Portals	3	709671	11.23	43.52	2	559254	3	456281	5	253390	1
Education	4	558183	8.84	52.36	5	304386	5	263847	3	294336	-1
Information Technology	5	449954	7.12	59.48	6	200185	6	181300	4	268654	-1
Web-based Applications	6	390278	6.18	65.66	4	336890	4	331990	11	58288	2
Games	7	379462	6.01	71.67	7	156351	7	145209	6	234253	0
Business	8	199455	3.16	74.83	9	108063	10	95567	7	103888	-1
Shopping	9	166820	2.64	77.47	11	94739	11	86591	8	80229	-2
File Sharing and Storage	10	163682	2.59	80.06	10	106536	9	102926	10	60756	0
Entertainment	11	153140	2.42	82.48	8	117183	8	113604	14	39536	3
Reference	12	152565	2.41	84.89	12	86090	12	78747	9	73818	0
Web-based Email	13	113965	1.8	86.69	13	68743	13	66548	12	47417	0
News and Media	14	99934	1.58	88.27	14	67278	14	65898	17	34036	0
Newsgroups and Message Boards	15	71043	1.12	89.39	16	35037	17	31629	15	39414	-1
Pornography	16	68720	1.09	90.48	15	42031	15	39897	18	28823	1
Personal Websites and Blogs	17	68312	1.08	91.56	20	25497	20	24055	13	44257	-3
Instant Messaging	18	62816	0.99	92.55	18	29973	18	28458	16	34358	0
Auction	19	55353	0.88	93.43	17	33344	16	32504	20	22849	2
Travel	20	48802	0.77	94.2	19	29955	19	24893	19	23909	1

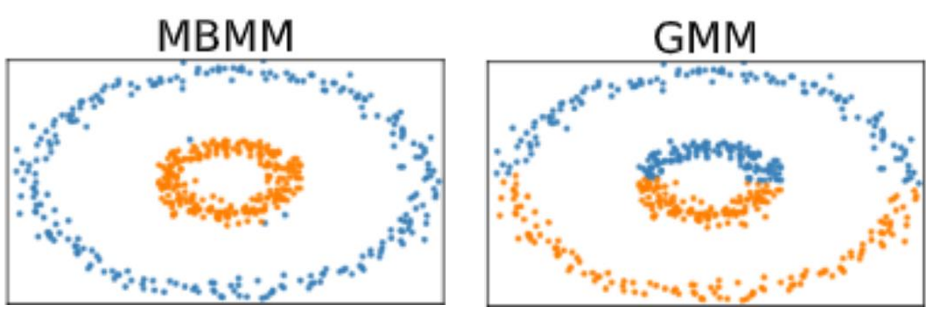
C.-Y. Hsu, T.-R. Chen, H.-H. Chen. **ACM Journal of Data and Information Quality** 14(2),

Multivariate Beta Mixture Model (MBMM) – ongoing

- A new probabilistic clustering algorithm
- Gaussian mixture model (GMM): each cluster has to be a Gaussian distribution
- MBMM: allow versatile shapes for each cluster
  - Uni-modal (symmetric or skewed), bi-modal



Versatile shape of bi-variate beta



MBMM vs GMM clustering

Y.-P. Hsu and H.-H. Chen. **PAKDD** 2024

Math Information Retrieval (MathIR) (1/2)

- Query: "$ax^2 + bx + c = 0$"
  - Does "$\alpha\theta^2 - \beta\theta = \gamma$" count as a match?
    - If x is the unknown in Eq1, θ is the unknown in Eq2, the two equations are equivalent
    - However, if compute similarity based on text-matching-based methods (e.g., TF-IDF), Eq1 and Eq2 are not similar
- Challenge
  1. MathIR is beyond text-matching
  2. Labeled data (i.e., which pairs of equations are similar) is limited

P.-S. Wang and H.-H. Chen. **IRonGraphs** 2024 (in conjunction with **ECIR** 2024)

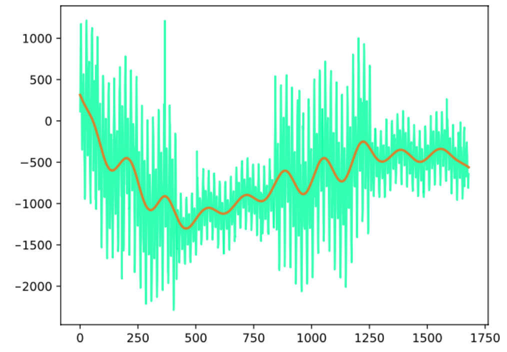
Math Information Retrieval (MathIR) (2/2)

- Convert equation into graphs
  - Capture the notation structures
    - Tackles challenge 1 (MathIR is beyond text matching)
- Graph Contrastive Learning (GCL)
  - Generate similar equation pairs based on graph augmentation techniques
    - Tacks challenge 2 (limited labeled data)
- Bpref scores (TangentCFT is a SOTA)

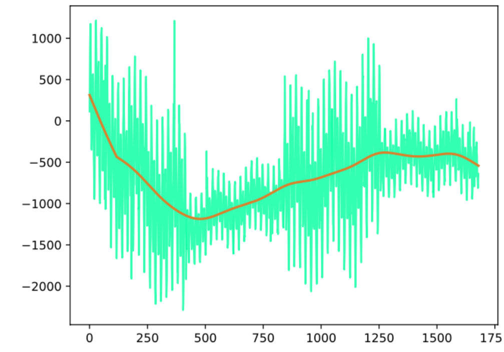
Model	SLT	OPT	F1
TangentCFT	0.680 ± 0.0053	0.660 ± 0.0064	0.670
InfoGraph	0.691 ± 0.0066	0.685 ± 0.0070	0.688
BGRL	$\mathbf{0.701} \pm 0.0089$	0.683 ± 0.0077	0.692
GraphCL	0.685 ± 0.0090	$\mathbf{0.703} \pm 0.0072$	**0.694**

P.-S. Wang and H.-H. Chen. **IRonGraphs** 2024 (in conjunction with **ECIR** 2024)

個人化之趨勢線生成 (1/2)



(a) Trend line 1



(b) Trend line 2

- 哪條才是趨勢線?
  - 不同情境，不同答案
  - 有人希望趨勢線「平滑」，有人希望趨勢線仍能有「局部起伏」
- 如何讓電腦「學習」一個人心中的趨勢線樣貌? -- 個人化趨勢線生成
- Training：系統展示十張時間序列，使用者分別標注其心目中的趨線，系統從中學習使用者想要的趨勢線樣貌
- Generation: 使用者給予系統所有需要標示趨勢線之時間序列，系統按 training 時學習到之規則自動為所有時間序列標出趨勢線
  - 挑戰：僅有十張訓練資料，如何有效的學習 (且不overfitting)

T.-Y. Kuo and H.-H. Chen, **PAKDD** 2023.

個人化之趨勢線生成 (2/2)

- 兩階段之 個人化趨勢線生成技術
- DNN model 容易 overfitting
- Pretrain and finetune有部份效果，但仍不理想
- Petrel (我們的方法) 優於上面兩類

Type	Algorithm	SMAPE	MSE
Our method	Petrel (averaged)	**0.44**	5264.34
	Petrel (weighted)	**0.44**	**5258.34**
DNN models	ConvNet	0.83	176593.87
	LSTM	1.02	497312.33
	Transformer	1.08	579188.89
DNN with pretraining and fine-tuning	P&F ConvNet	**0.44**	5425.77
	P&F LSTM	0.52	7394.09
	P&F Transformer	0.47	9311.75
	P&F MLP	0.68	31934.92

資料集一

Algorithm	SMAPE	MSE
Petrel (averaged)	0.33	6164.38
Petrel (weighted)	**0.32**	**6002.32**
ConvNet	0.94	166951.8
LSTM	1.11	323712.95
Transformer	1.20	637955.96
P&F ConvNet	1.45	241890.91
P&F LSTM	1.23	1292454.44
P&F Transformer	0.81	1357013.58
P&F MLP	1.18	242234.14

資料集二

T.-Y. Kuo and H.-H. Chen, **PAKDD** 2023.

E-commerce object and behavior embedding (Behavior2Vec)

- Predict a user's next clicked item
- Predict a user's next purchased item
- Discover the relationship between items
  - E.g., Canon's camera body : Canon's lens ≈ Nikon's camera body : Nikon's lens

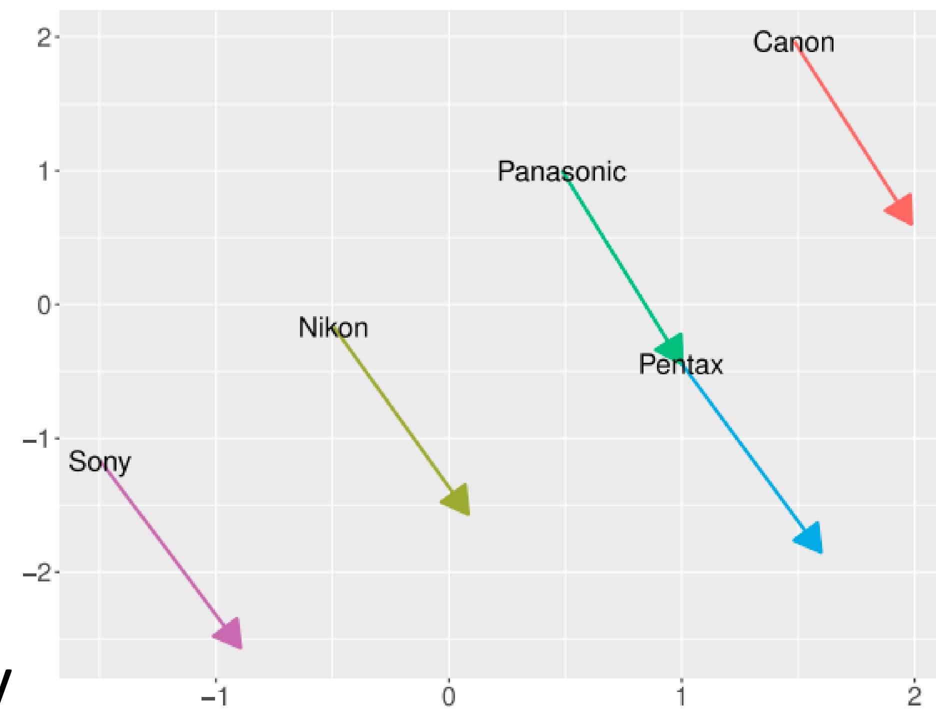


Figure 1: Vectors from the camera body to the corresponding kit lens of different brands. The vectors are generated by Behavior2Vec

H.-H. Chen. **ACM TKDD** 2018.

Recommendation for near cold start items

- Near cold start item: items that are rarely viewed
- Recommendation for the near cold start items is difficult because of the limited clues
- Our RDF method alleviates this issue

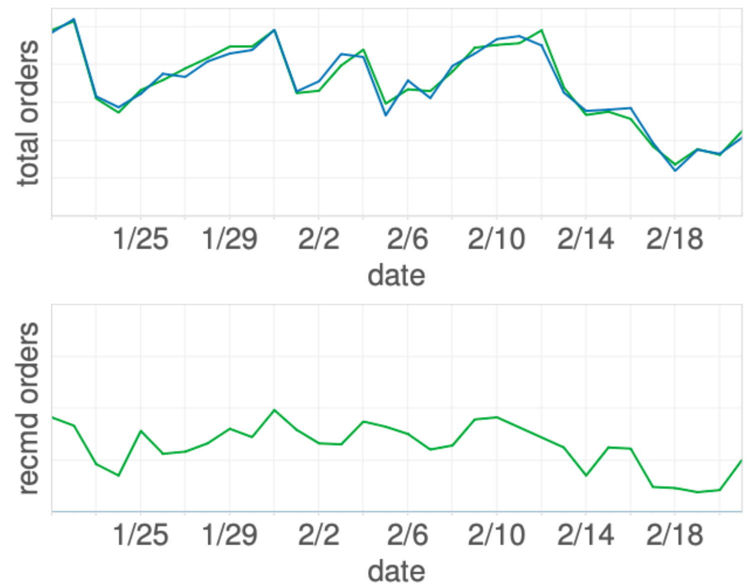
Table 1: a comparison of the methods with RDF and without RDF

Dataset	SVD	linear-reg	sqrt-reg	log-reg	improve ratio range
Epinions	1.1997	**1.0538**	**1.0538**	**1.0538**	12.16%
MovieLens-100K	0.9423	**0.9422**	**0.9422**	**0.9422**	0.01%
FilmTrust	0.8465	**0.8194**	**0.8194**	0.8223	2.86% to 3.20%
Yahoo! Movies	3.0799	**2.9892**	3.0129	3.0127	2.18% to 2.94%
AMI	1.1450	**1.1405**	**1.1405**	**1.1405**	0.39%

H.-H. Chen and P. Chen. **ACM TKDD** 2019.

Train and evaluate recommender systems in the right way

- Show 4 common errors in training and evaluating recommender systems
- Propose solutions or work-arounds for these issues



Green: channel with a recommendation
Blue: channel w/o recommendation

H.-H. Chen, C.-A. Chung, H.-C. Huang, W. Tsui. **ACM SIGKDD Explorations** 2017.

Co-learning user's browsing tendency of multiple categories

- Instead of predicting each target variable independently, our MFMT method simultaneously learns multiple targets in one model

Table: F1 scores of different models on different target categories

model	shopping	traveling	restaurant and dining	entertainment	games	education
kNN	0.574	**0.615**	0.528	0.440	0.492	0.484
Logreg	0.578	0.489	0.501	0.402	0.441	0.437
SVM	0.576	0.391	0.410	0.399	0.409	0.385
MFMT	**0.584**	0.570	**0.561**	**0.479**	**0.531**	**0.515**
	(win)		(win)	(win)	(win)	(win)

G.-J. Bai, C.-Y. Lien, H.-H. Chen. IEEE/WIC/ACM International Conference on **Web Intelligence** 2019

User personality and demographic profile prediction based on browsing logs

Table: errors of the personality test score prediction based on the supervised learners with and without the preprocessing step

Method	Supervised regressor						Clustering + supervised regressor (win)					
Prediction target	HH	Neu	Ext	Agr	Con	Ope	HH	Neu	Ext	Agr	Con	Ope
Lasso	5.832	5.87	5.881	5.71	5.406	5.607	**5.411**	**5.469**	**5.435**	**5.435**	**5.022**	**5.131**
Ridge	5.845	5.981	5.891	5.795	5.43	5.646	**5.43**	**5.404**	**5.38**	**5.325**	**5.027**	**5.052**
Elastic net	5.813	5.769	5.743	5.622	5.366	5.44	**5.417**	**5.383**	**5.422**	**5.317**	**5.022**	**5.095**
SVR	5.789	5.78	5.746	5.643	5.232	5.38	**5.432**	**5.623**	**5.402**	**5.328**	**5.048**	**5.165**

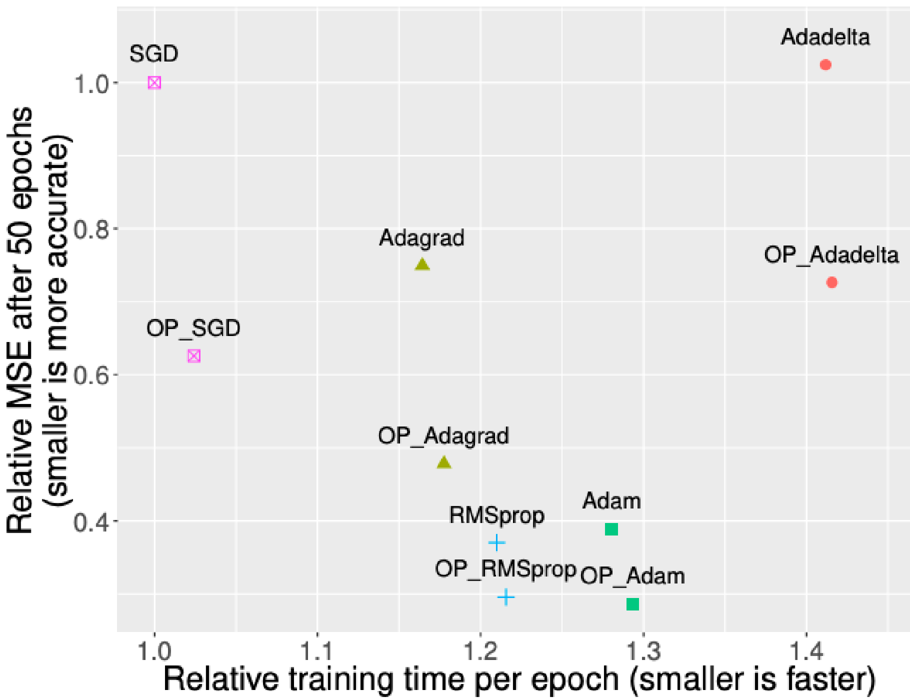
Table: MicroF1 scores of the demographical info prediction based on the supervised learners with and without the preprocessing step

Method	Supervised classifier			Clustering + supervised classifier (mostly win)		
Prediction target	Age	Gender	Relationship	Age	Gender	Relationship
Baseline	0.388	0.545	0.474	**0.411**	**0.598**	**0.476**
KNN	0.427	0.594	0.478	**0.435**	**0.618**	**0.482**
Random Forest	**0.453**	**0.697**	0.488	0.419	0.687	**0.512**
Logistic Regression	0.427	**0.697**	0.476	**0.457**	0.675	**0.498**
SVM	0.388	0.591	0.474	**0.411**	**0.642**	**0.512**

C.-Y. Lien, G.-J. Bai, H.-H. Chen. IEEE/WIC/ACM International Conference on **Web Intelligence** 2019.

Accelerating MF by Overparameterization

- Overparameterization significantly accelerates the optimization of MF
  - Theoretically derive that applying the vanilla SGD on OP_MF is equivalent to using GD with momentum and adaptive learning rate on the standard MF model



P. Chen and H.-H. Chen. **DeLTA** 2020.

Public transportation optimization

- Predict the taxi demand in real time by deep learning

Model	RMSE	MAPE
Average	8.845 ± 7.9434	0.0840 ± 0.000413
ARIMA	15.585 ± 20.8253	0.1660 ± 0.018033
ridge regression	10.914 ± 2.4451	0.1460 ± 0.000895
XGBoost	6.498 ± 2.0542	0.0806 ± 0.000205
LSTM (2 layers)	7.037 ± 3.9747	$\boxed{0.0563 \pm 0.000056}$
LSTM (4 layers)	6.694 ± 5.1110	0.0595 ± 0.000232
DMVST-Net	7.350 ± 3.7034	0.0643 ± 0.000192
ResLSTM (4 layers)	$\boxed{5.187 \pm 2.0265}$	0.0584 ± 0.000048
AR-LSTM (4 layers)	$\boxed{4.958 \pm 1.8909}$	$\boxed{0.0488 \pm 0.000039}$

C.-J. Hsu and H.-H. Chen. **VEHITS** 2020.

台灣大車隊載客熱點
預測競賽**第二名**

Dynamic ensembled learning

- Dynamically integrate multiple base learners based on the feature distribution of the test instance
- Better accuracy than the static ensembled learning approach

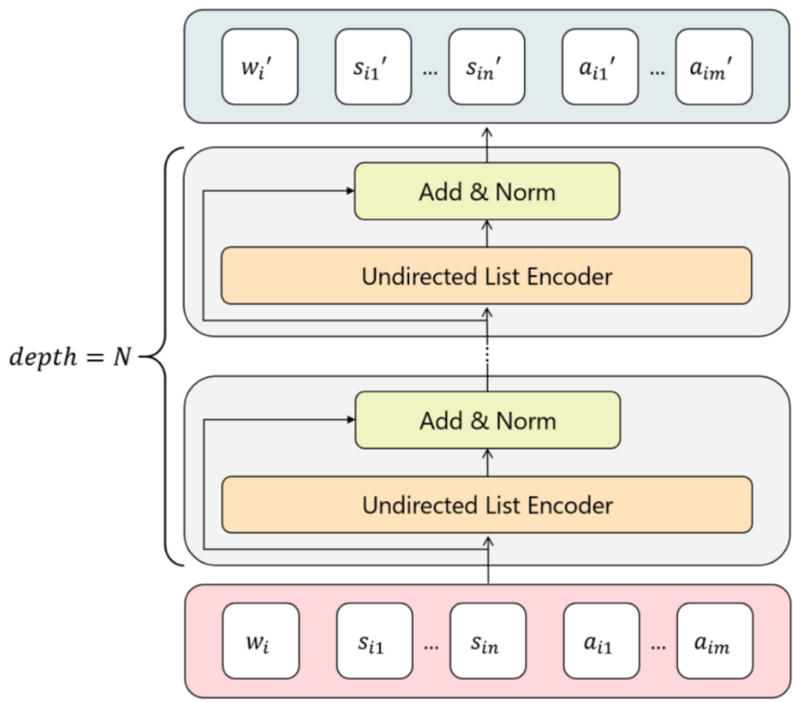
Table: a comparison of the base learners, static ensembled, and dynamic ensembled methods

Method	KNN	SVM	Decision Tree	Majority Voting	Dynamic ensembled
Accuracy	77.09%	72.77%	75.46%	77.64%	77.80% (win)

蘇俊儒, 陳弘軒. 動態多模型融合分析研究. **TANET** 2018. (最佳論文獎)

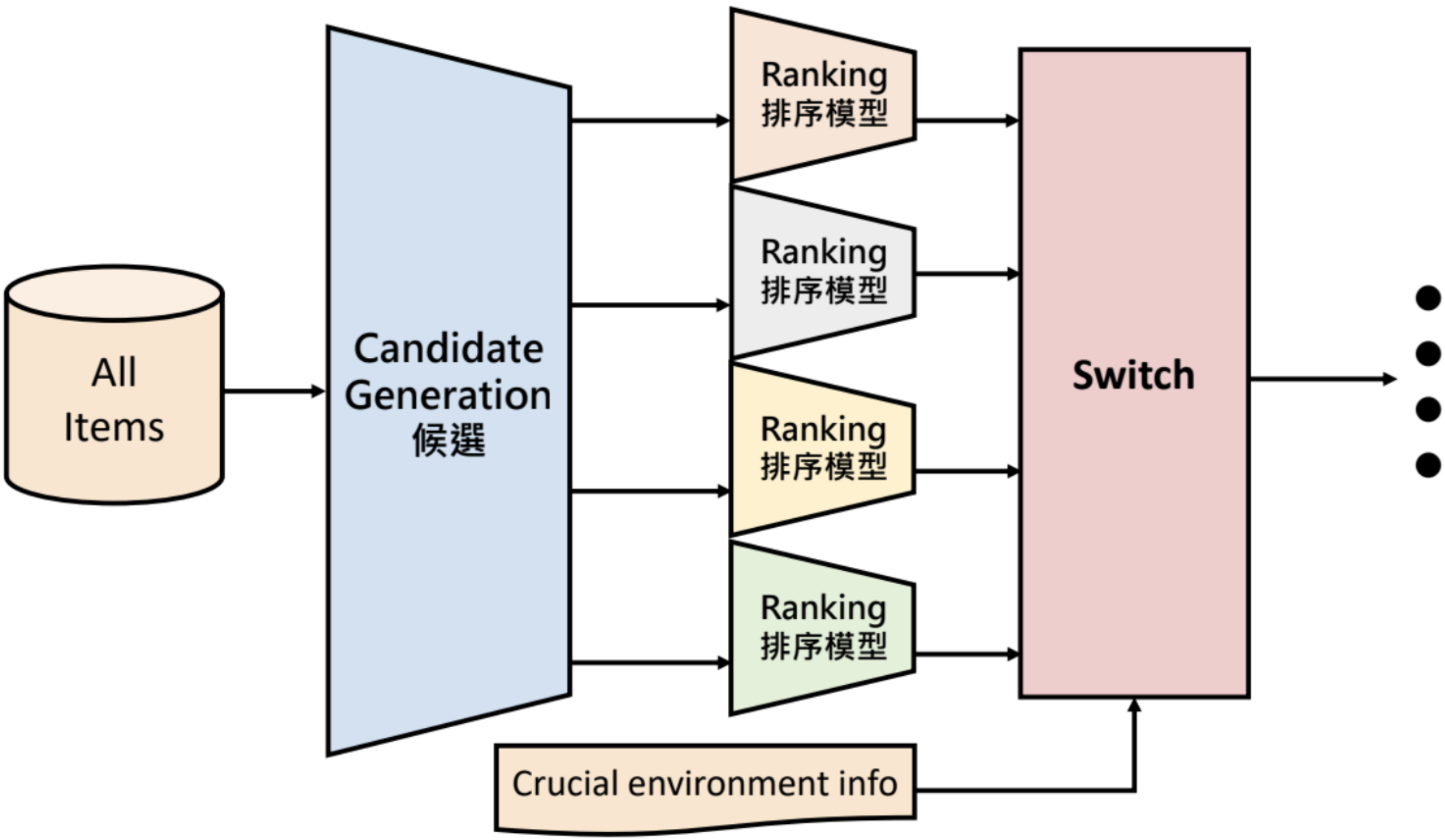
Better word embedding for synonyms and antonyms

- Adjusting word embedding to differentiate synonyms and antonyms



林冠佑，中大論文 2020

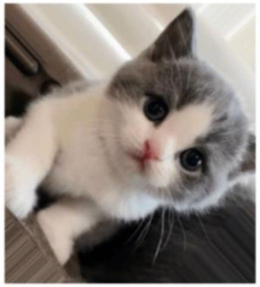
Deep vs shallow recommendation

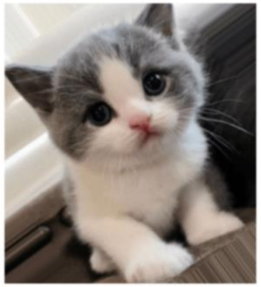


Y.-C. Yang, P.-C. Lai, H.-H. Chen. **TAAI** 2020.

Recent undergraduate projects (大學專題)

抗旋轉之卷積網路設計





- 卷積網路難以判斷旋轉的圖片
  - 從電腦的角度來看，旋轉前後的圖片之pixel排列方式不同，故可能認定圖片中為不同的物件
- 深度學習通常需要讓電腦看過各種旋轉角度的圖片，讓電腦「認得」不同旋轉角度的相同物件
- 我們設計新的模型，電腦只需看過一張圖，即可認得各種旋轉角度的圖片

Test accuracy

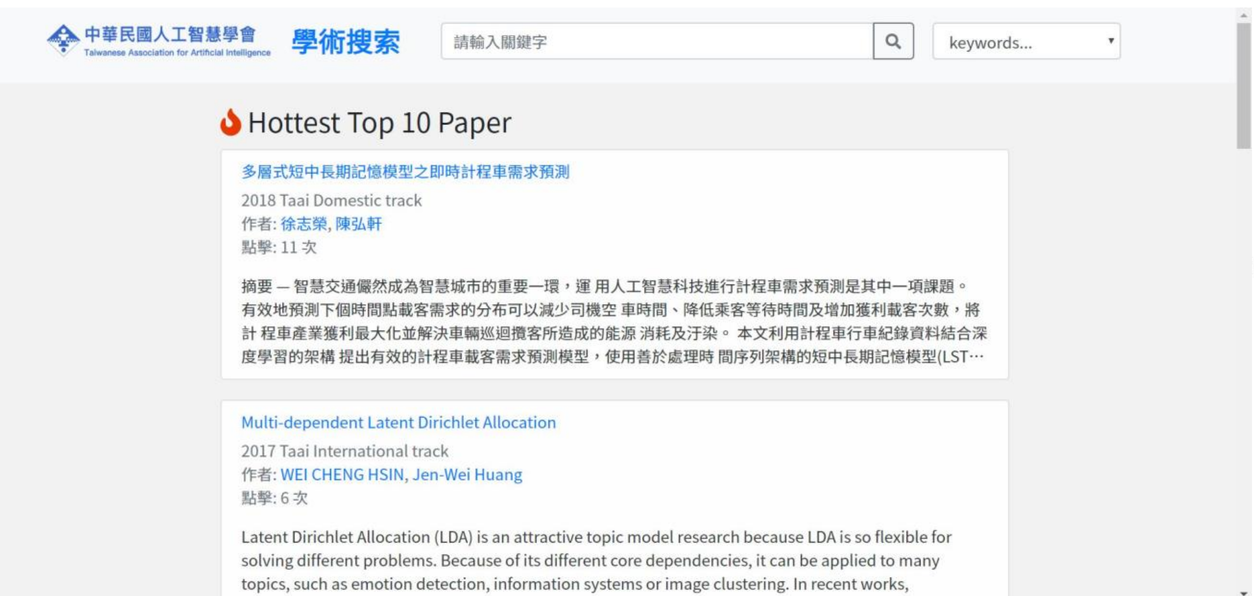
	MNIST		FashionMNIST		CIFAR-10	
	轉90度	任意旋轉	轉90度	任意旋轉	轉90度	任意旋轉
ConvNet1	0.17	0.42	0.07	0.22	0.29	**0.30**
ConvNet2	0.16	0.33	0.02	0.19	0.24	0.23
Our model	**0.72**	**0.43**	**0.79**	**0.33**	**0.36**	0.26

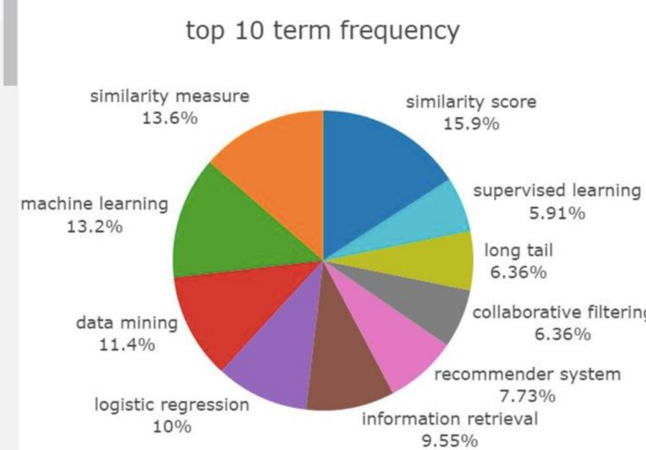
楊佳誠、陳弘軒. **TAAI** 2022

2022 資工系專題競賽**最佳創意獎**

學術搜尋引擎/關鍵字標註器

- Build an academic search engine for the Taiwanese Associationfor Artificial Integlligence (中華民國人工智慧學會)
  - http://search.taai.org.tw/
- Keyword search
- Paper keyphrase extractor



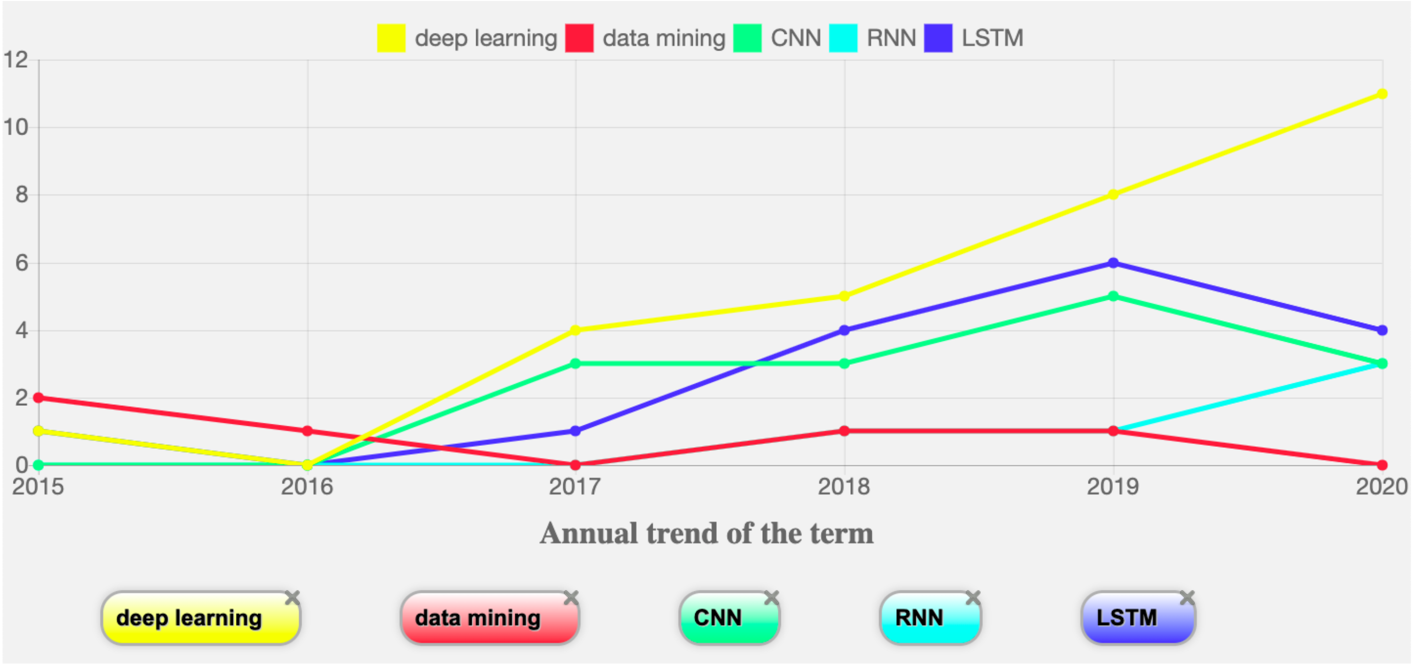


李竺芸、林冠佑、陳弘軒、張嘉惠. **TAAI** 2020

2020 資工系專題競賽**第一名**

基於學術搜尋引擎之研究趨勢分析

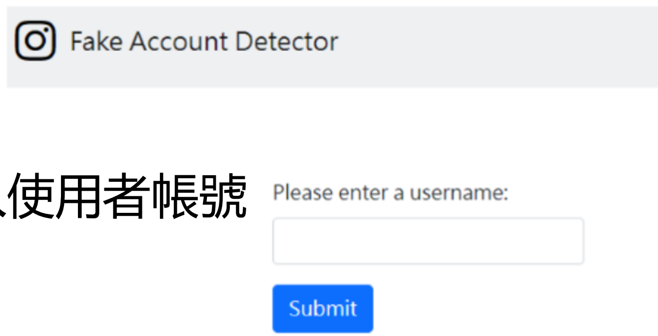
- 利用 TAAI 論文發表內容分析研究主題歷年趨勢



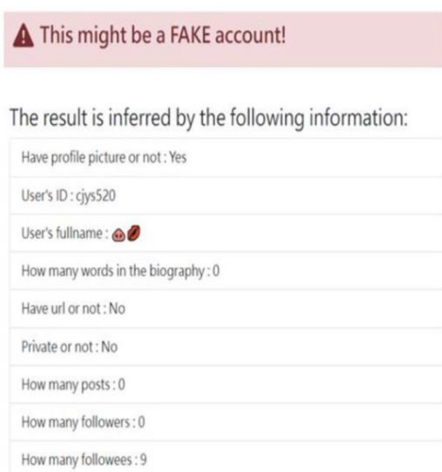
作者：林冠甫、王承隆

Instagram騷擾帳號偵測

初始畫面

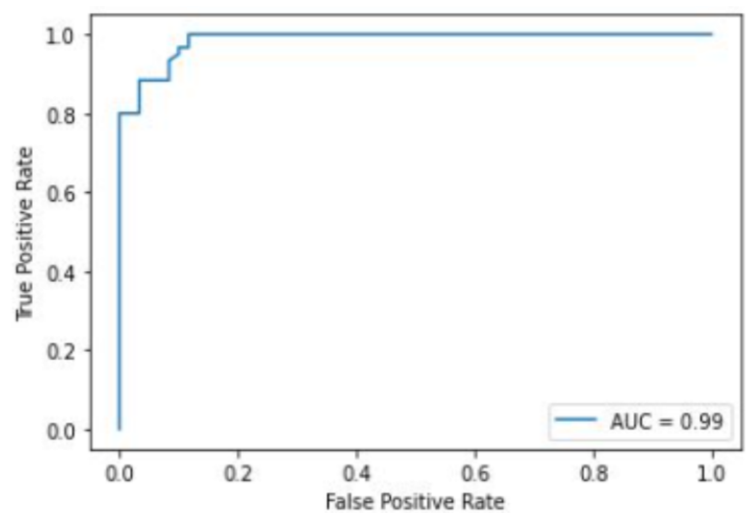


輸入使用者帳號



自動判斷該帳號是否為騷擾帳號

- 模型效能
  - Accuracy: 0.925
  - Precision: 0.932
  - Recall: 0.917
  - AUROC: 0.99



作者：陳映璇, 楊沂潔

精彩影片片段自動截取

- Generate highlight clips and thumbnails for videos based on bullet-screen (彈幕) information
- Our model is better than the software used by video streaming companies

Table 1: Users' evaluation on the representativeness of the outputted video clips

	All users	Group 1	Group 2
Busk	**52.12%**	**67.38%**	47.45%
Stiller	47.88%	32.62%	**52.55%**

Table 2: Users' evaluation on the representativeness of the outputted thumbnail images

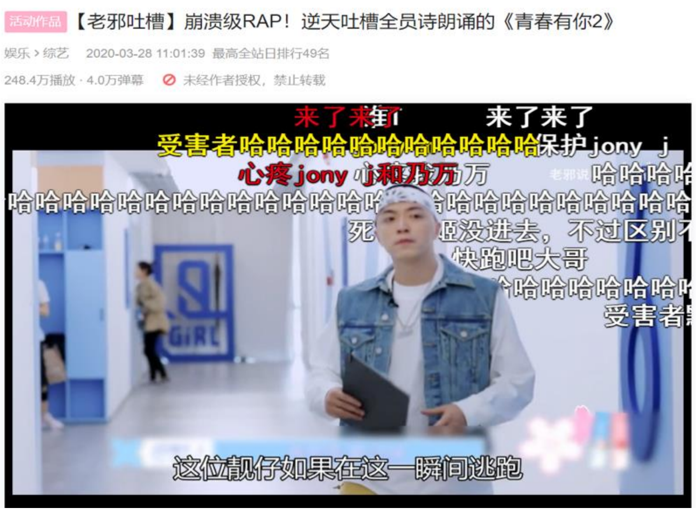
	All users	Group 1	Group 2
Busk	47.62%	**63.08%**	43.39%
Stiller	**52.38%**	36.92%	**56.61%**

- Group 1: users who are familiar with the videos
- Group 2: others

Yun-Ya Huang (黃筠雅), Tong-Yi Kuo (郭同益), Hung-Hsuan Chen (陳弘軒).
WWW 2020, **TAAI** 2019

影片「片段」搜尋

- Search for the video clips inside long videos based on bullet-screen (彈幕) information



作者：呂晨瑀、王心玓

2020 資工系專題競賽**最佳人氣獎**

從 FAQ 自動產生 Chatbot

- 從常見問題集 (FAQ) 自動產生客服對話機器人
- 對話機器人利用 Elasticsearch、Word2Vec、及 BERT 判斷「使用者的問題」與「常見問題集中各問題」的相似度
- 對話機器人回傳相似的常見問題及答案

重設密碼

Top 3 matches:
(1) 我是畢業生，忘記密碼，無學生證認證身分，該如何修改密碼?
(2) 我的帳號仍未失效，要如何更改密碼?
(3) 我是教職員身分，忘記密碼該如何處理?

周冠玲、歐亭昀、陳弘軒. **TAAI** 2021　　2021 資工系專題競賽**入選獎**

PDF 數學式解析器

- 要讓電腦 “瞭解” 文件中的數學式，第一步需要讓電腦能解析數學式
  - E.g., $(a+b)^2 \Rightarrow$ (a + b)^2
- PDF是科學論文最常見的格式
- 為了在不同裝置能有一樣的文件外觀，PDF 描述每個符號應該以怎樣的型式 (e.g., 大小、字型、顏色等) 出現在哪個位置
  - 這使得數學式很難被自動化的解析
- 我們採機器學習+自訂規則解析PDF中的數學式

$$ax^2 + bx + c = 0$$

$$e^{i\pi} + 1 = 0$$

PDF 文件

數學式解析器

ax^2 + bx + c = 0
e^{i\pi} + 1 = 0

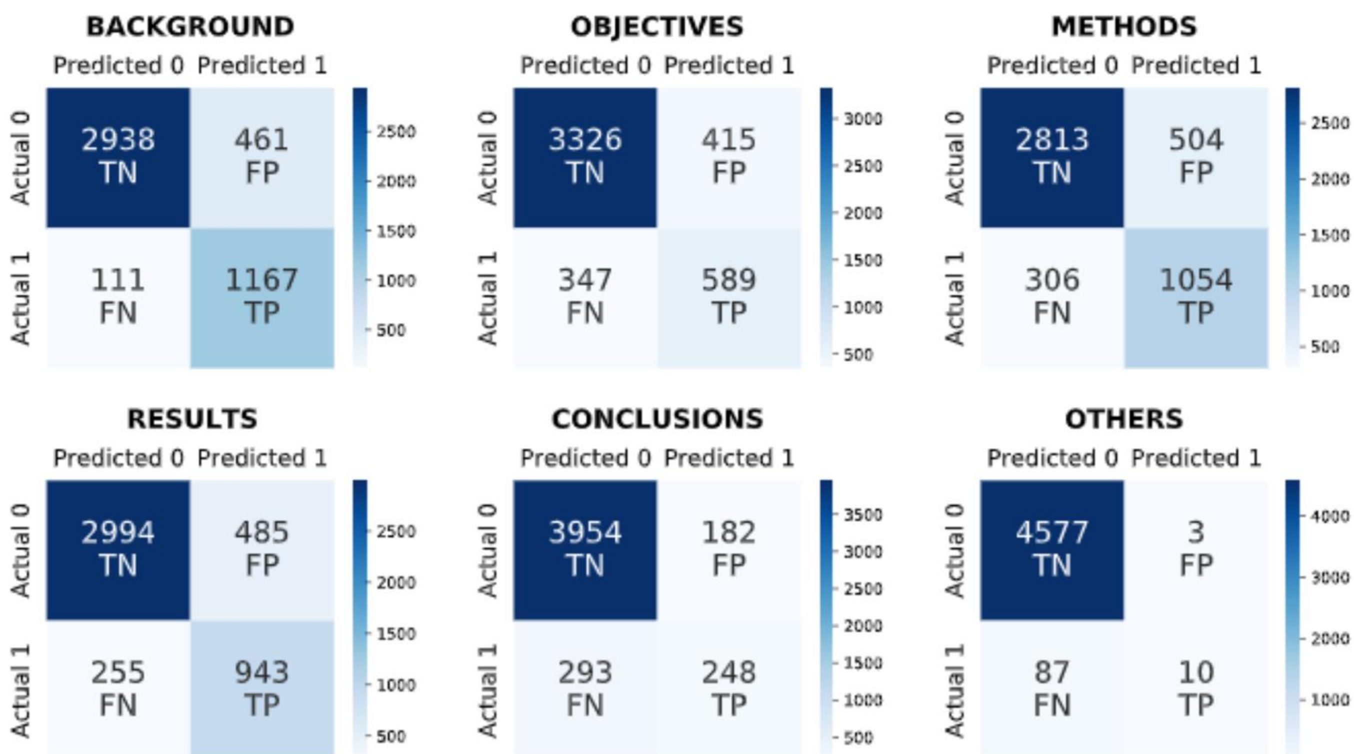
LaTeX 格式

作者：謝欣玉

2021 資工系專題競賽**入選獎**

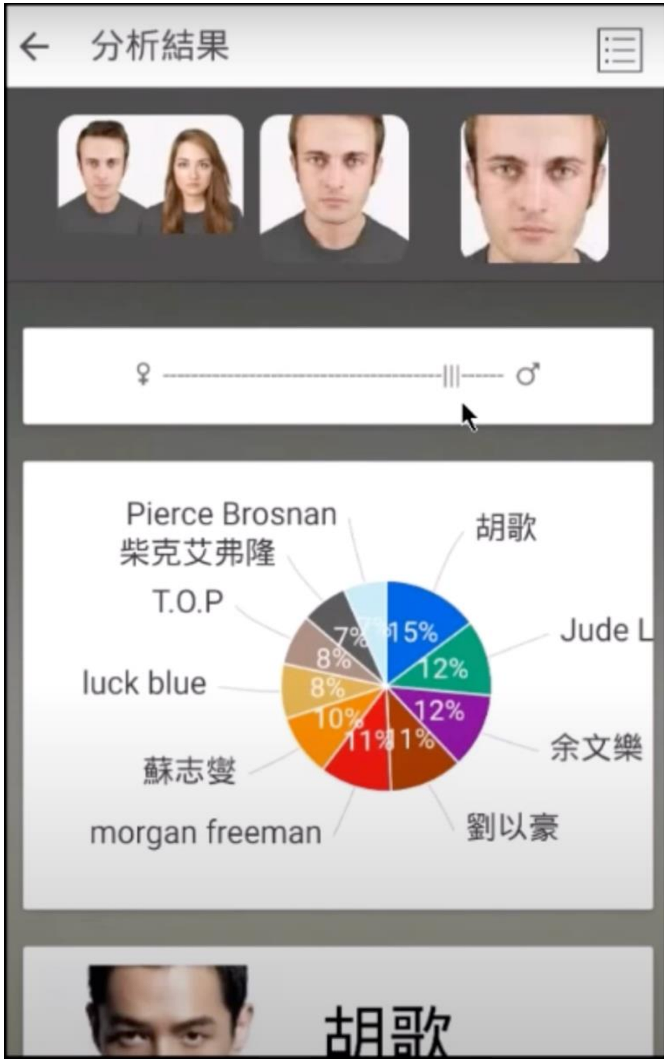
論文摘要的句子之撰寫目的預測

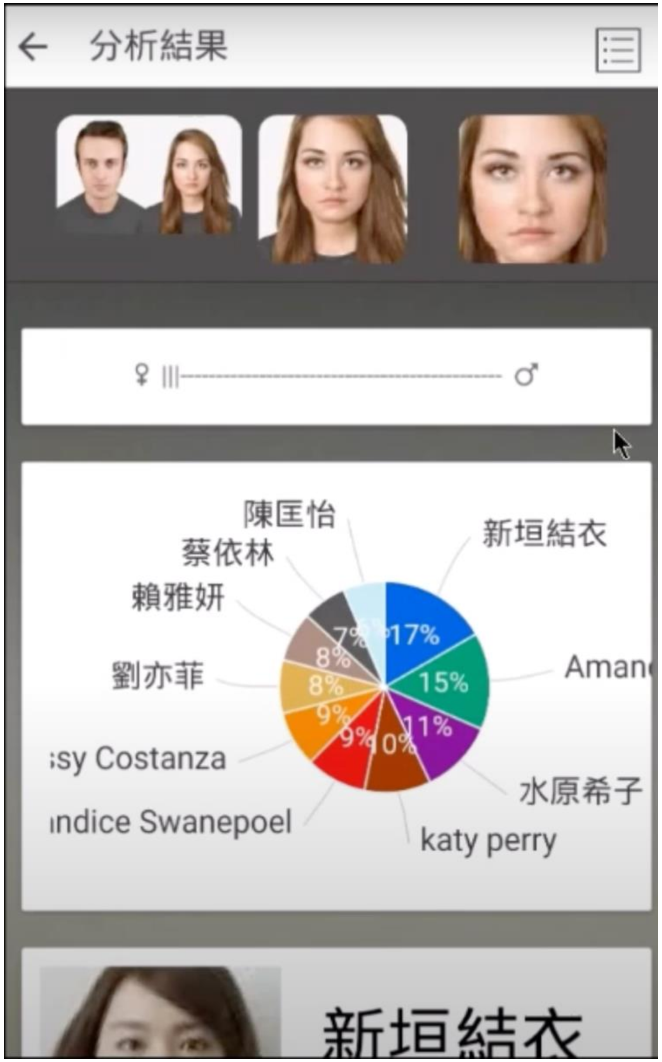
- Predict the “purposes” of each sentence in an abstract



作者：許力元、高嘉豪、鄭易昇、陳弘軒. **IEEE Access** 2021

我是大明星－明星臉分析





作者：蔡文傑

行動應用服務APP競賽**第三名**