INCOMPLETE LABEL UNCERTAINTY ESTIMATION FOR PETITION VICTORY PREDICTION WITH DYNAMIC FEATURES

Junxiang Wang, Yuyang Gao, Andreas Züfle, Jingyuan Yang and Liang Zhao

International Conference on Data Mining (ICDM) 2018

GEORGE MASON UNIVERSITY

Content

- Introduction
- Problem Formulation
- Challenges
- Our Method
- Optimization
- Experimental Results

Introduction: Background

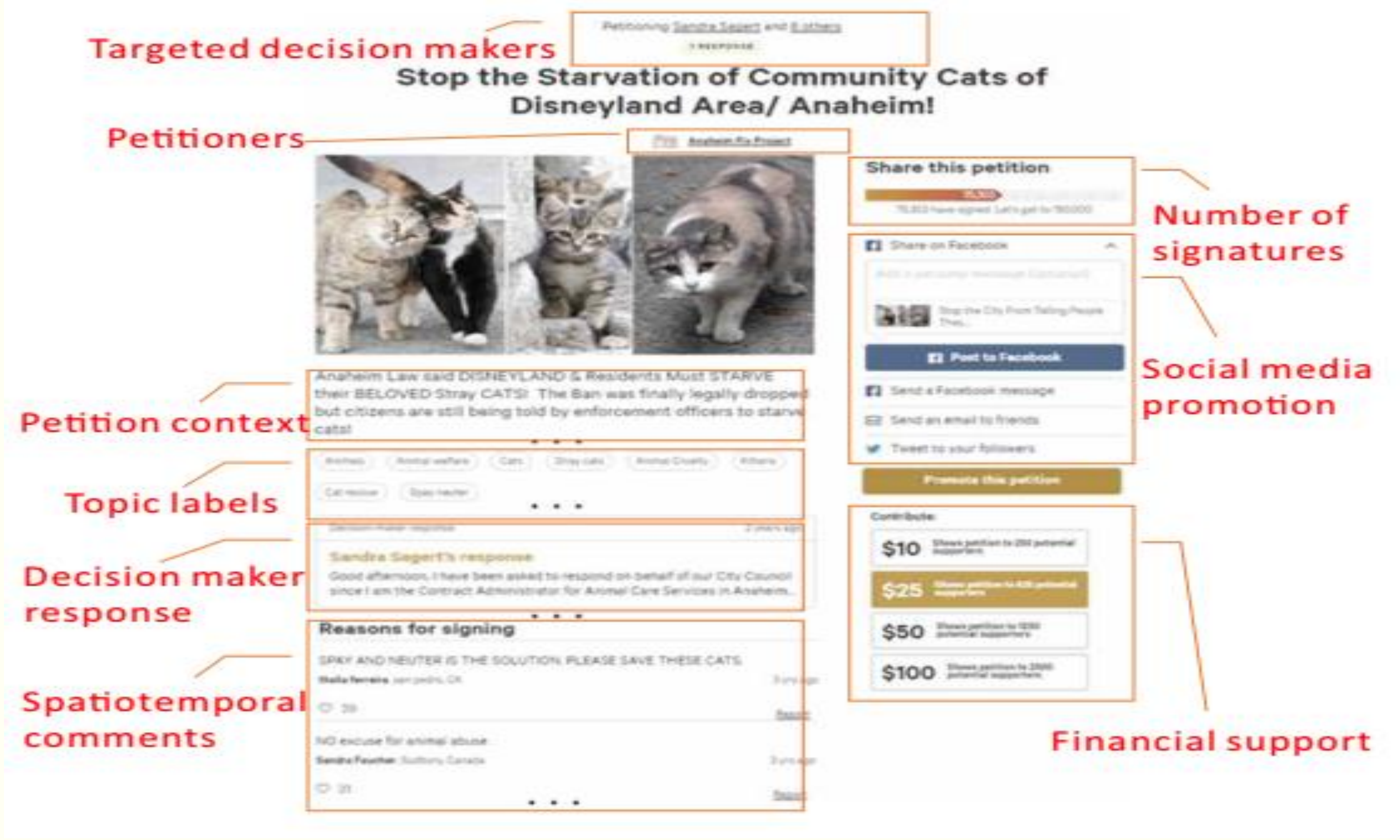
- The rise of Online Petition Platform (OPP) spurred with the internet and social networking.

➢ Change.org, which was founded in 2007, has owned over 190 million users and hundreds of daily petitions covering various social aspects by July 2017.

- These petition websites can fill the gap between the increasing public concerns and the decision-makers' attention.

➢ By making the decision-makers of Whole Foods Market agree on stopping discarding "ugly looking fruits", online petitions which aimed at reducing serious food waste were achieved.

Introduction: An example of Online Petitions



Introduction: Pros and Cons of Online Petitions

- An online petition is easily created by using web-based hosts to gain enough signatures in order to gain the attention of responsible decision-makers.

<table>
<tr><th>Pros</th><th>Cons</th></tr>
<tr><td>Low financial cost</td><td rowspan="2">poorly organized due to
1: massive nearly-duplicated and correlated petitions
2: sophisticated problems of spatio-temporal and semantic dissemination from various similar petitions.</td></tr>
<tr><td>Easy accessibility</td></tr>
</table>

- Motivation to this work: prioritize petitions with higher victory probability more efficiently and proactively

Problem Formulation

- A petition will be labeled as victorious if

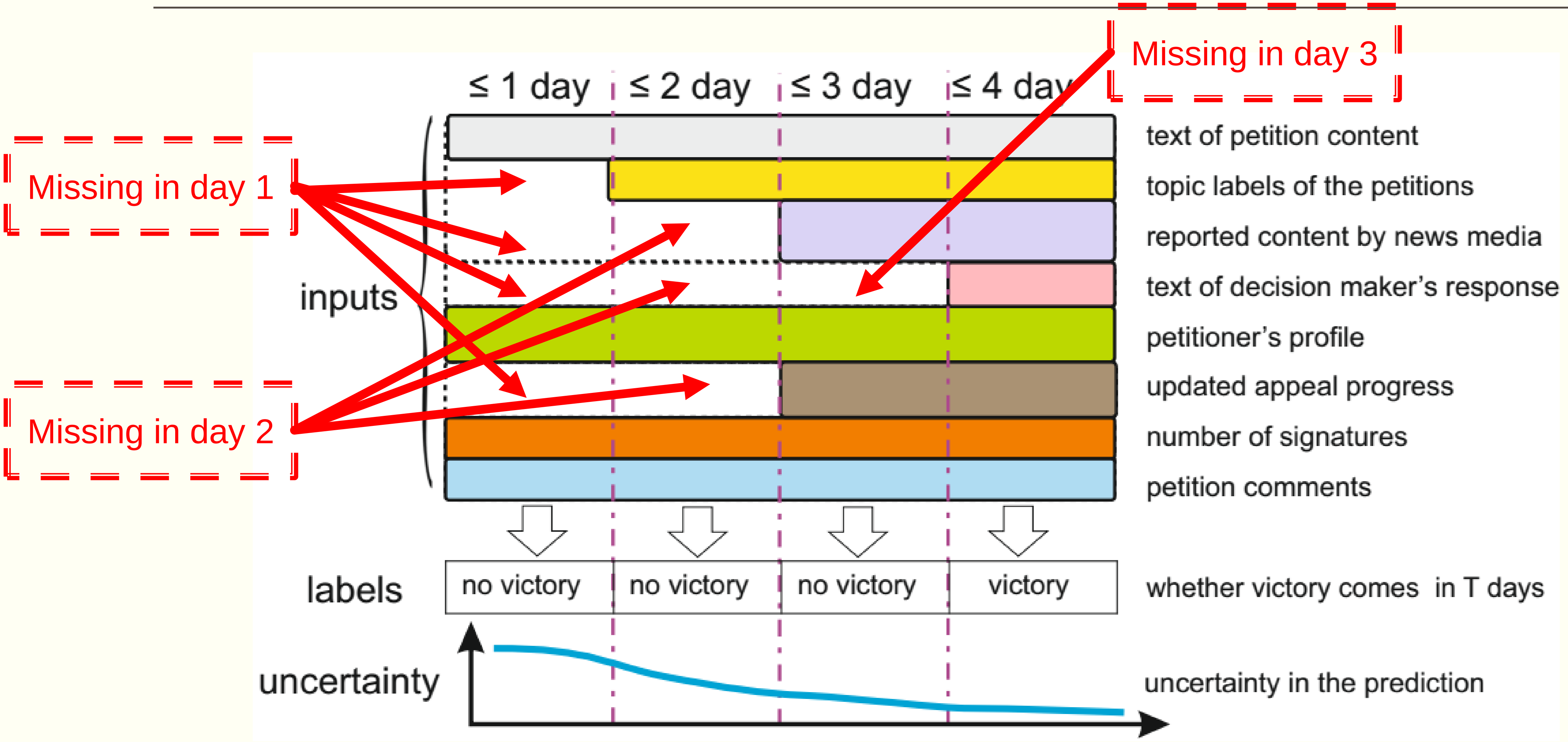
➢ (1). The required number of signatures is satisfied or

➢ (2). The appeals of the petition launcher have been addressed by the decision-makers within a limited time interval.

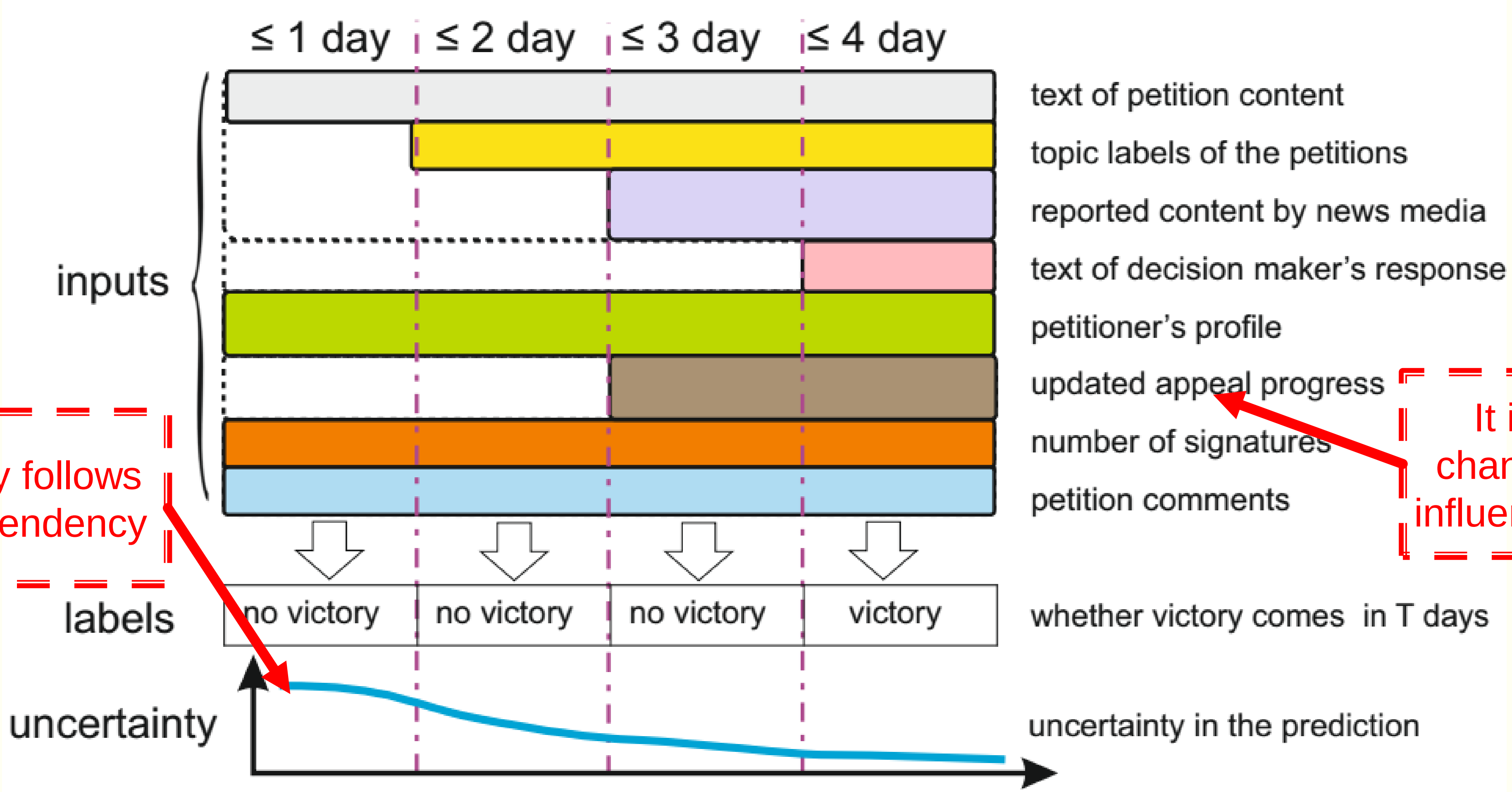
- The petition victory prediction problem is formulated as

➢ Given the petition vector $X_{i,t}$, the goal of this problem is to predict whether the i-th petition will succeed at time $t+\tau$ by learning the mapping $f: X_{i,t} \rightarrow Y_{i,t+\tau}$, where τ is the lead time.

Challenges: 1. Missing values in dynamic features



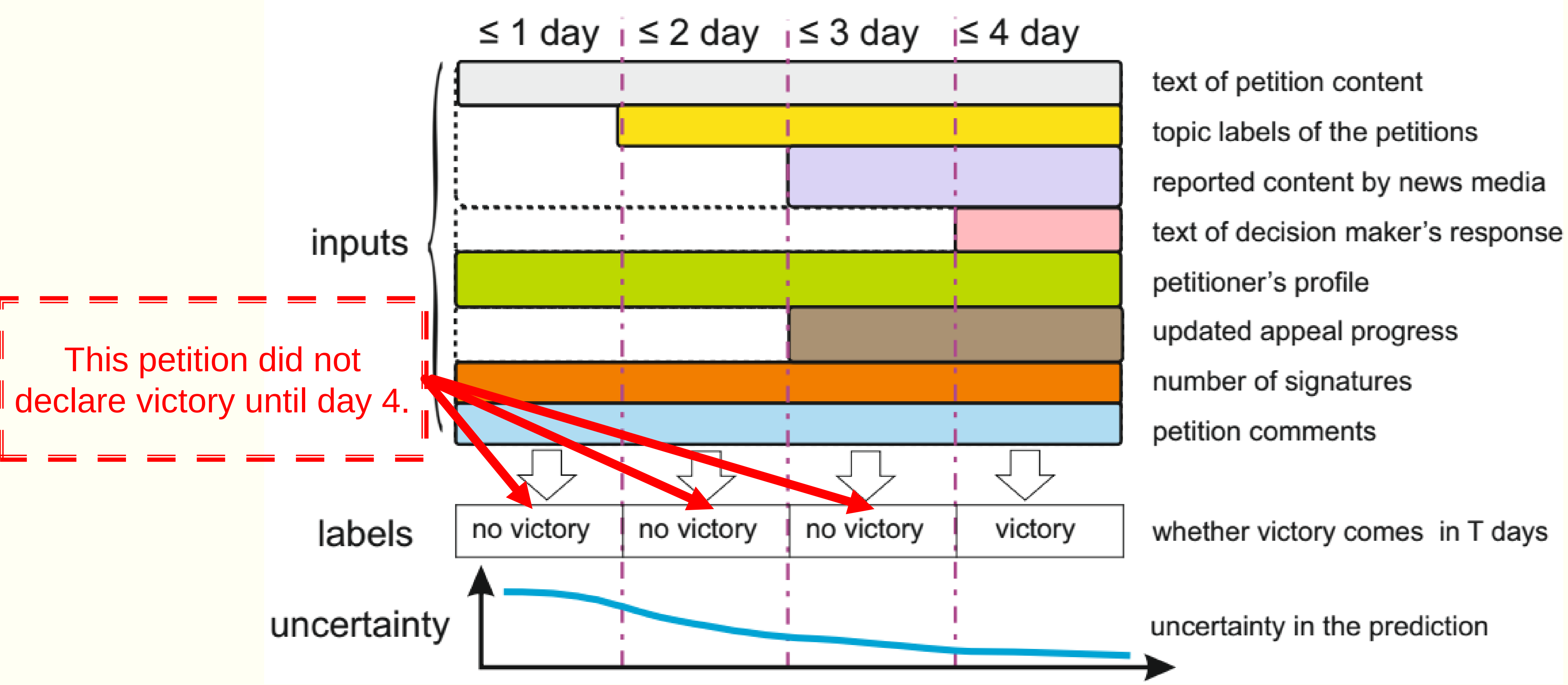
Challenges: 2. Strong uncertainty in petition prediction



The uncertainty follows a temporal dependency

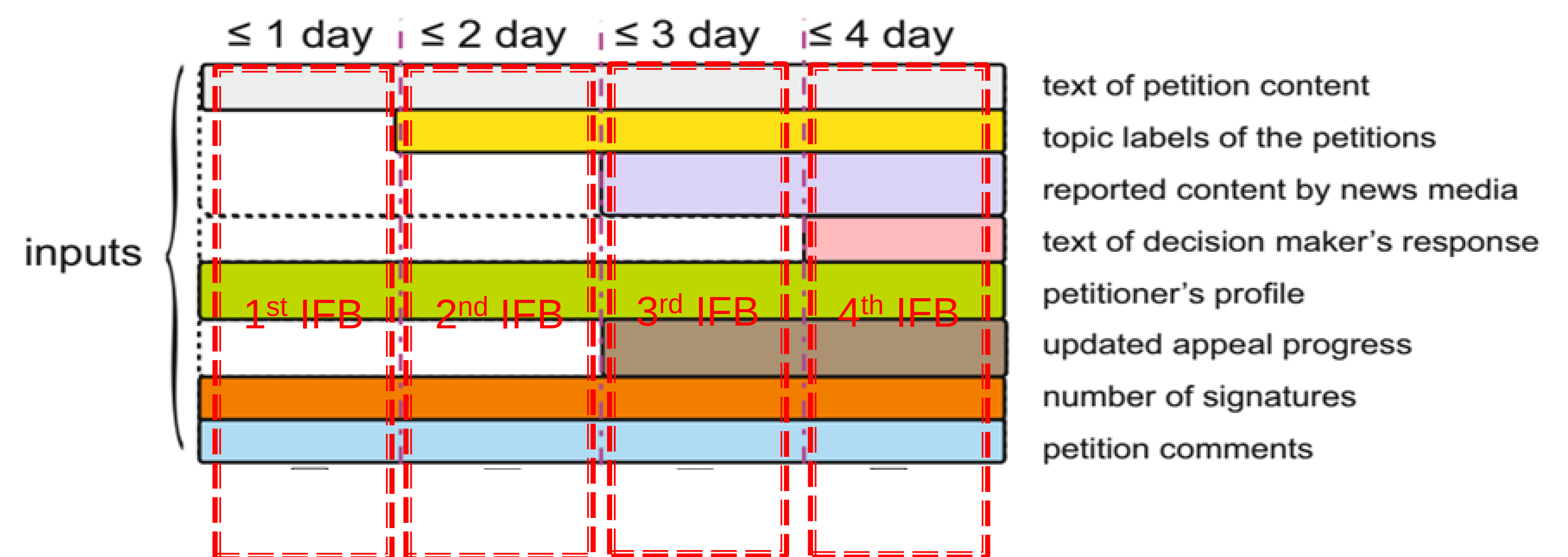
It is subject to change, which will influence uncertainty.

Challenges: 3. Unknown labels for ongoing petitions



Our Method: Multi-task Learning model with Uncertainty Estimation(MLUE)

- Increasing Feature Block(IFB)
  - ➢ Completeness. The time intervals partitioned by IFBs are complete.
  - ➢ Coherence. Petition sets in the same IFB share the same missing patterns.
  - ➢ Orderliness. The available features in the j-th IFB is contained in the (j+1)-th IFB.
- We consider each block as an independent task.



Our Method: Multi-task Learning model with Uncertainty Estimation(MLUE)

▪ Uncertainty Estimation

- If the predicted label in j-th IFB is correct while that in (j-1)-th IFB is wrong, we **earn more certainty.**

- If the predicted label in j-th IFB is wrong while that in (j-1)-th IFB is correct, we **lose more certainty**.

the accuracy earning of the classifier

the accuracy losing of the classifier

$$\begin{aligned} earn(Y_{i,q}, Y_{i,p}, Y_{i,d_i}) &= I(Y_{i,q} \neq Y_{i,d_i}) I(Y_{i,p} = Y_{i,d_i}) \\ &= (1 - Y_{i,q} Y_{i,d_i})(1 + Y_{i,p} Y_{i,d_i})/4 \\ lose(Y_{i,q}, Y_{i,p}, Y_{i,d_i}) &= I(Y_{i,q} = Y_{i,d_i}) I(Y_{i,p} \neq Y_{i,d_i}) \\ &= (1 + Y_{i,q} Y_{i,d_i})(1 - Y_{i,p} Y_{i,d_i})/4 \end{aligned}$$

uncertainty function

$$R(Y_{i,q}, Y_{i,p}, Y_{i,d_i}) = lose(Y_{i,q}, Y_{i,p}, Y_{i,d_i}) - earn(Y_{i,q}, Y_{i,p}, Y_{i,d_i})$$

i: the i-th petition.
d_i: the labeled time.
$Y_{i,p} (p \in T_j)$: the predict label in IFB(j).
$Y_{i,q} (q \in T_{j-1})$: the predict label in IFB(j-1).
Y_{i,d_i}: the label of the i-th petition at time d_i.

Our Method: Multi-task Learning model with Uncertainty Estimation(MLUE)

- Overall model

Empirical loss Regularization term Uncertainty function

$$(Y, \beta^*, b^*) = \arg\min_{Y,\beta,b} \boxed{Loss(Y;\beta,b)} + \boxed{\lambda_1 \Omega(\beta)} + \boxed{\lambda_2 R(Y)}$$
$$\boxed{s.t. \forall q \le p, Y_{i,q} \le Y_{i,p}} \quad (1)$$

Non-decreasing order

β: the set of coefficients of all tasks.
b: the set of intercepts of all tasks.
Y: the set of all petition labels at time intervals(It contains known labels and unknown labels).

Optimization: Expectation-Maximization(EM)-like algorithm

- This objective is nonconvex because Y is discrete. Therefore, we propose an EM-like algorithm to solve it.

➢ E-step: update Y when fixing β and b.

➢ M-step: update β and b when fixing Y.

- Updating Y: dynamic programming.
- Updating β and b: Alternating Direction Method of Multipliers(ADMM).

Experimental Results: Dataset

- Petition data: Two-stage data collection.

➢ The first stage: we queried the Change.org API to obtain information from 54,039 petitions during Jan 1, 2009 and Dec 17, 2017.

➢ The second stage: all corresponding comments were retrieved by Change.org API again.

Petition Field	Number of Fields
basic properties	10
petition topic	21
petition tag	128
petition title	288
petition description	711
petition body	210
victory description	79
petition comments	74
all	1521

Country	Victorious #Petitions	Failed #Petitions	Ongoing #Petitions
Philippines	60	202	750
India	237	2,527	3,110
German	776	2,691	5,594
Australia	479	1,374	2,525
Canada	398	1,475	1,951
United States	4,081	7,405	18,404

Experimental Results: Metrics and comparison methods

	Predicted failure	Predicted victory
Actual failure	TN	FP
Actual victory	FN	TP

- Metrics:

➢ Accuracy(ACC)=(TN+TP)/(TN+FP+FN+TP)

➢ Precision(PR)=TP/(FP+TP)

➢ Recall(RE)=TP/(FN+TP)

➢ F-score(FS)=2RE*PR/(PR+RE)

➢ Area Under ROC curve(AUC): compute the Area under Receiver Operating Characteristic (ROC) curve.

- Comparison methods:

➢ Multi-task learning: Constrained Multi-Task Feature Learning I (cMTFL-I), convex relaxed Clustered Multi-Task Learning(CMTL), multi-task learning with Joint Feature Selection (JFS).

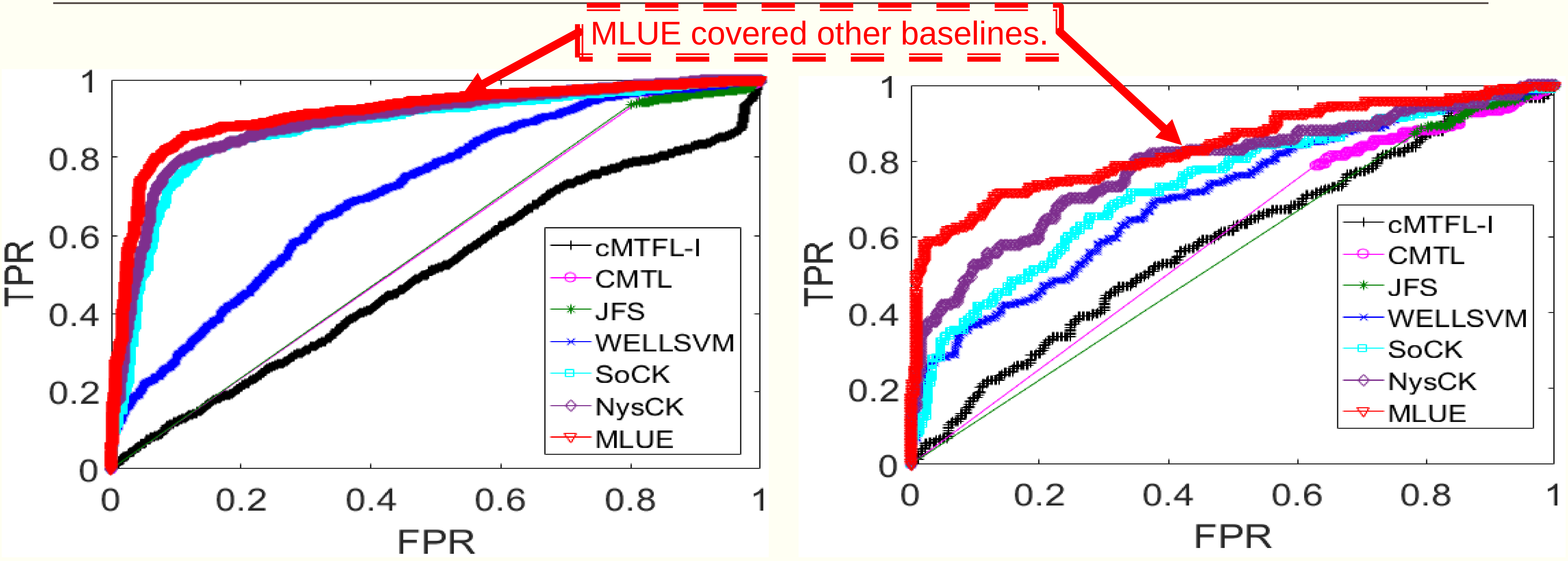
➢ Semi-supervised learning: WEakly LabeLed Support Vector Machines (WELLSVM), Stochastic optimization for Cluster Kernel(SoCK), Nystrom Cluster Kernel (NysCK).

Experimental Results: Petition victory prediction on US dataset

	Methods	ACC	PR	RE	FS	AUC
Multi-task learning	cMTFL-I	0.5435	0.5588	0.9377	0.7003	0.4869
	CMTL	0.5674	0.5695	**0.9799**	0.7203	0.5842
	JFS	0.5735	0.5738	0.9773	0.7228	0.5996
Semi-supervised learning	WELLSVM	0.5359	0.7648	0.4915	0.4451	0.7250
	SoCK	0.8202	0.8644	0.8112	0.8369	0.8719
	NysCK	0.8347	0.8882	0.8117	0.8482	0.8880
	MLUE	**0.8602**	**0.9247**	0.8212	**0.8698**	**0.9140**

- MLUE ranked the first in four metrics out of five.
- Semi-supervised learning methods outperformed multi-task learning methods.

Experimental Results: ROC curve on US dataset and Canada dataset

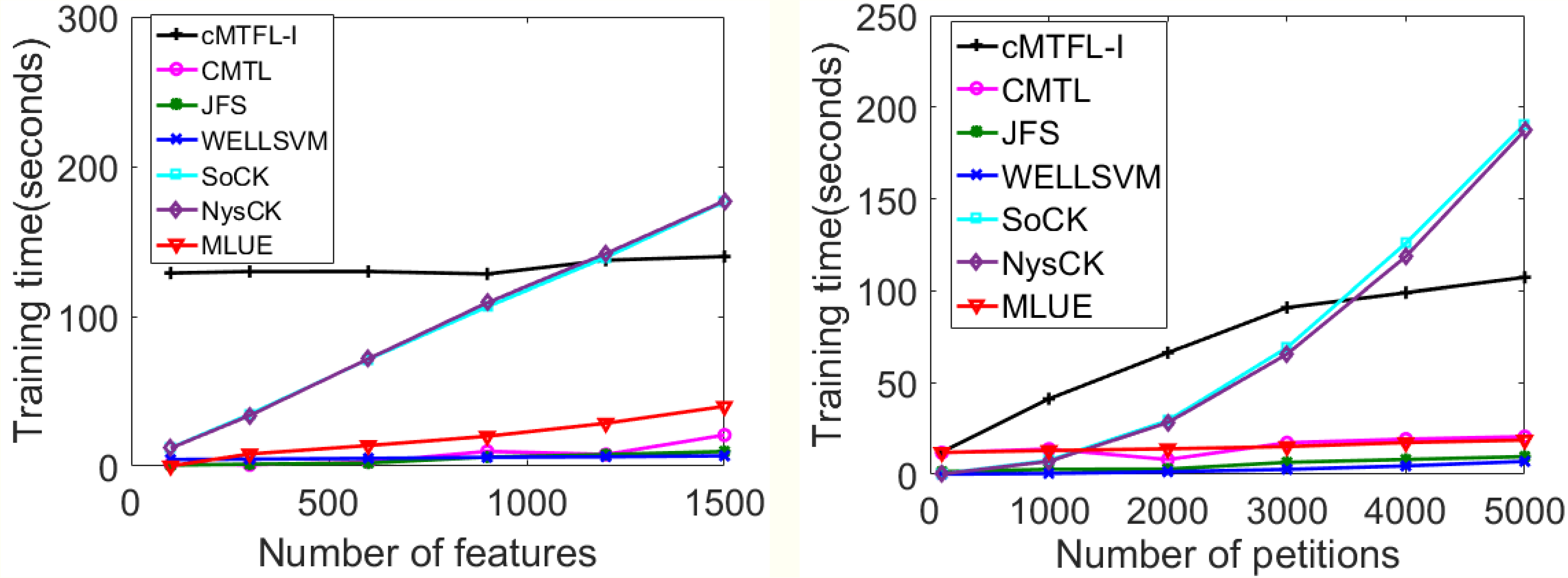


US dataset

Canada dataset

- MLUE performed the best in general.
- Semi-supervised learning methods outperformed multi-task learning methods.

Experimental Results: Scalability analysis



- Nearly all methods increase linearly with number of features and petitions.

Experimental Results: Feature analysis on US dataset

Education -related features

Right - related features

Petition Field	Feature	Weight Value
victory description	department	3.3964
petition tag	k-12	3.0348
petition tag	healthcare	2.9356
petition tag	gay rights	2.7908
petition tag	disability rights	2.7803
petition tag	students	2.7134
petition tag	clemency	2.6915
petition tag	education	2.6883
victory description	continue	2.6612
petition tag	cats	2.5314

Our Dataset

The link of our dataset and code

http://mason.gmu.edu/~lzhao9/materials/data/petition/index.html

Feel free to contact me (jwang40@gmu.edu) if you have any questions.

Thank you. Any questions?