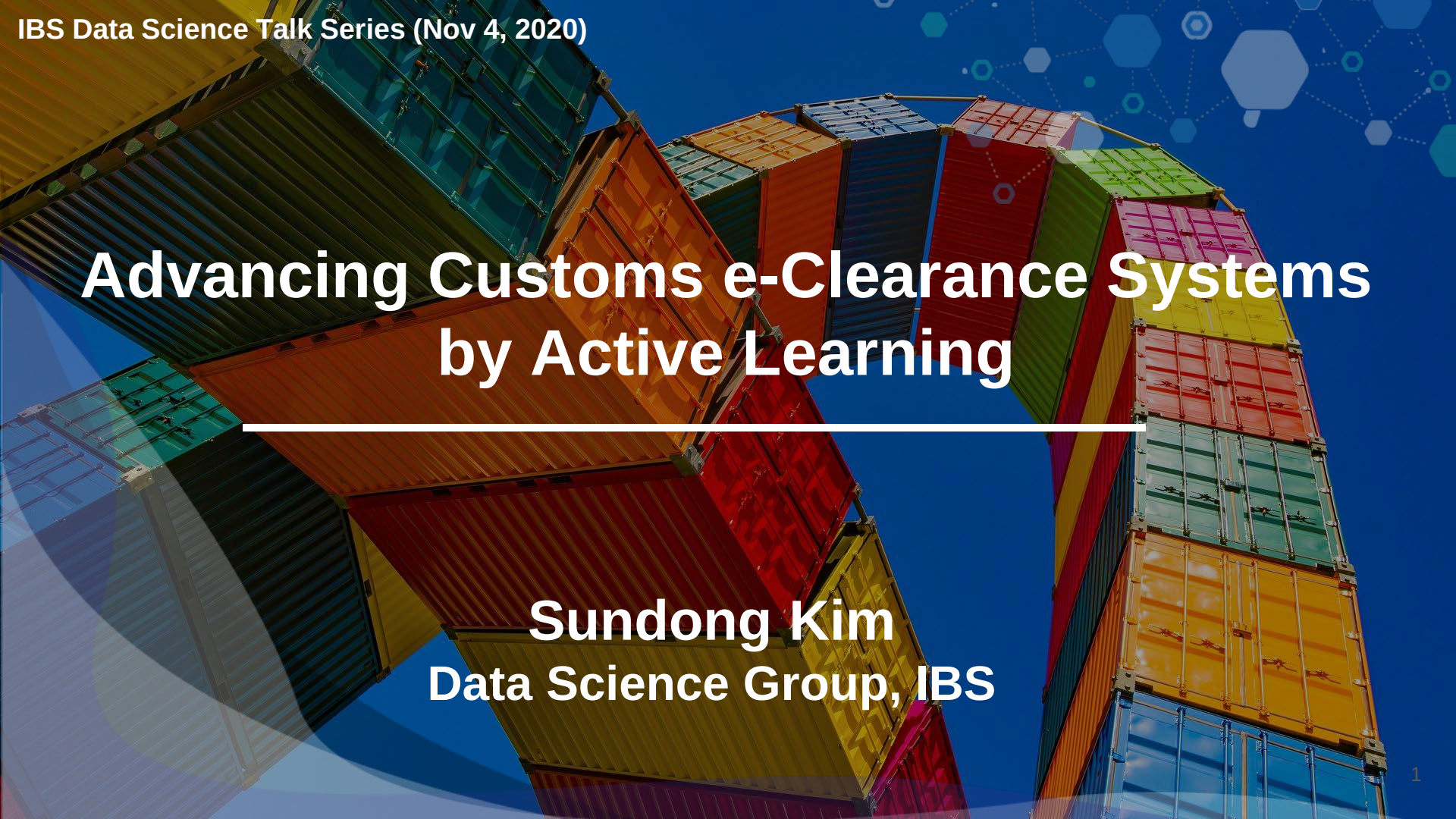


Advancing Customs e-Clearance Systems by Active Learning

Sundong Kim
Data Science Group, IBS

Supporting customs with data analytics

Other articles in this Edition >>

Flash Info

BACUDA: supporting Customs with data analytics

By the WCO Secretariat

WCO Members asked the Secretariat to place a new focus on the development of guidance and capacity to support the use of data analytics. As one of the responses, a team of experts was put in place under a project called BACUDA. The project's name is an acronym, which stands for "BAnd of CUstoms Data Analysts." It is also a Korean word that means "to change." Indeed, the aim of the project is to help Customs administrations in embracing analytical tools and methodologies, a major move for many.

BACUDA team members are all data experts with whom the Secretariat has been collaborating for some years. They are Customs officials in charge of risk management, statistics and IT systems, as well as professional economists and data scientists with an academic background in computer science. Data scientists of various nationalities from the Institute of Basic Science (IBS), the Korea Advanced Institute of Science and Technology (KAIST), and the National Cheng Kung University (NCKU) are involved in the project and leading the development of state-of-the-art algorithms. However, any qualified data experts working in Customs administrations or in academia may join the BACUDA team.

NIGERIA CUSTOMS SERVICE

KOREA CUSTOMS SERVICE 1878

http://bit.ly/IBS-BACUDA-news
https://ds.ibs.re.kr/index.php/bacuda/

Our project

“AI-based system for targeting customs inspection”

Can we develop a safe and fast machine learning model for customs e-clearance system

Type of Frauds	Illicit motives
Undervaluation of trade goods	To avoid ad-valorem customs duty, or conceal illicit financial flows from exporters
Misclassification of HS code	To get a lower tariff rate applied or trade prohibited goods by avoiding restriction
Manipulation of origin country	To get a preferential tariff rate under a free trade agreement

Targeting High Risk Records
FNN & CNN (KCS+KISTI, 2018)
RF using Risky Profiles (WCJ, 2019)
SVM Ensembles (Belgium, 2020)
DATE (IBS+WCO+NCKU, 2020)

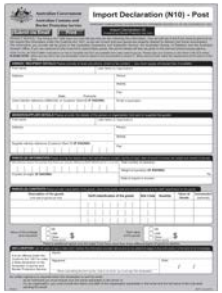
FYI: Import clearance procedure

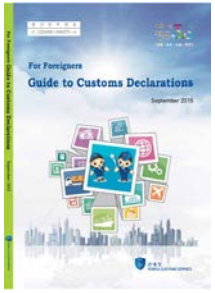
Showing how to declare imports into Uni-pass:
1) https://www.youtube.com/watch?v=WE1gkH-VucE
2) https://www.youtube.com/watch?v=0AWATNYHnx4
3) https://www.youtube.com/watch?v=nQjdNMVf4pU

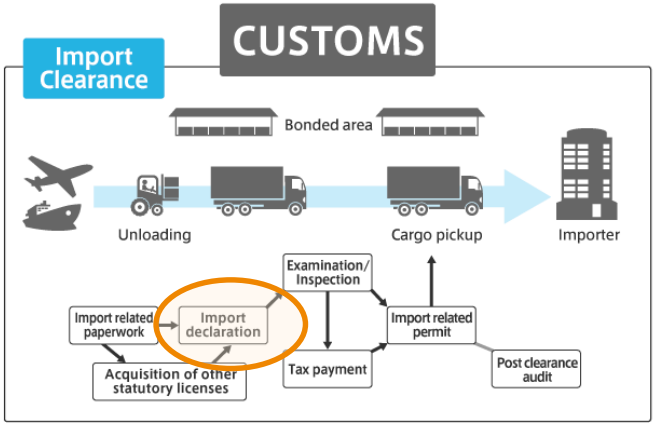
1. Prior to import, an importer or a forwarder make a pre-declaration of the item.
2. Once your cargo enters Korea, you will receive disembarkation message.
3. The cargo is moved to the warehouse declared by the forwarder in advance.
4. When there is no problem, the cargo is taken into the warehouse.
5. When your cargo is brought into the warehouse, forwarder can start its import declaration process.
6. After filing a declaration, there would a screening process.
   (Targeting system - Customs selection operates here) If the item is subject to inspection, customs officers will check the document or have physical inspection.
7. If it passes without being specified, a payment notification step appears.
8. Import declaration process will be completed once the customs duty is paid, and you can take out your items out of the warehouse.

* forwarder = declarant

Import declarations



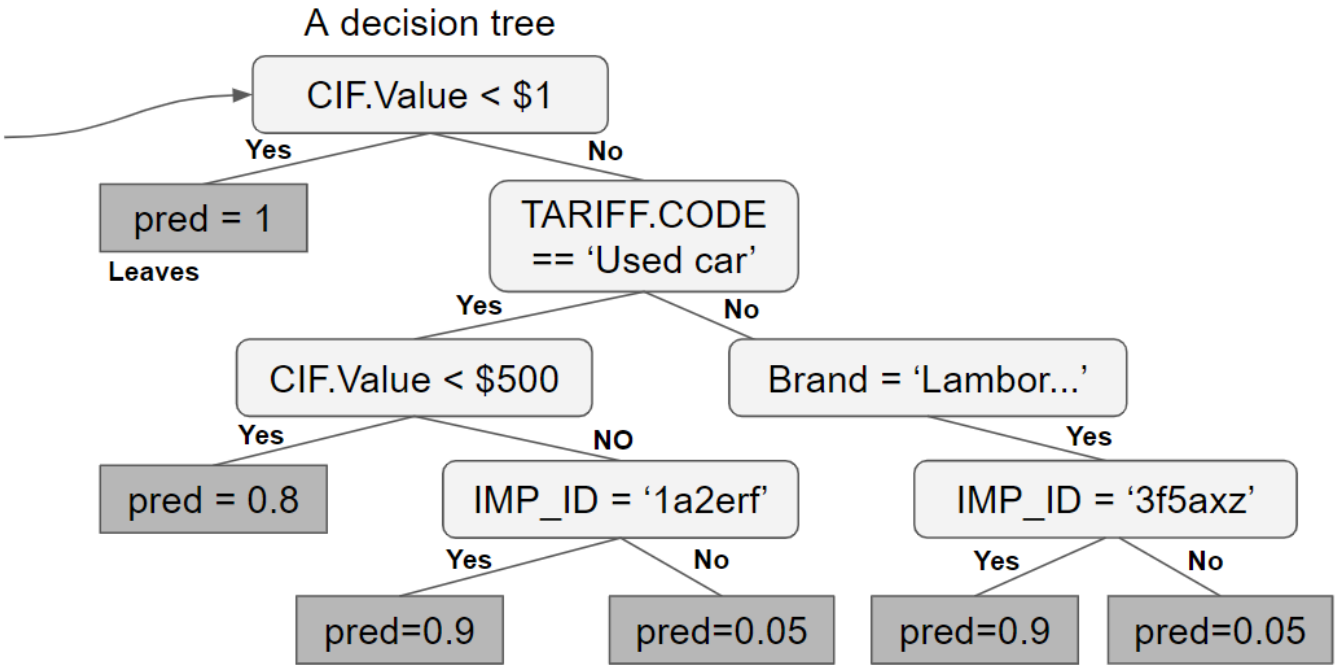




Type	Variable	Description	Example
Features	*sgd.id*	An individual numeric identifier for Single Goods Declaration (SGD).	SGD347276
	sgd.date	The year, month and day on which the transaction occurred.	13-11-28
	importer.id	An individual identifier by importer based on the tax identifier number (TIN) system.	IMP364856
	declarant.id	An individual identification number issued by Customs to brokers.	DEC795367
	country	Three-digit country ISO code corresponding to transaction.	USA
	office.id	The customs office where the transaction was processed.	OFFICE91
	tariff.code	A 10-digit code indicating the applicable tariff of the item based on the harmonised system (HS).	8703232926
	quantity	The specified number of items.	1
	gross.weight	The physical weight of the goods.	150kg
	fob.value	The value of the transaction excluding, insurance and freight costs.	$350
	cif.value	The value of the transaction including the insurance and freight costs.	$400
	total.taxes	Tariffs calculated by initial declaration.	$50
Prediction Target	***illicit***	Binary target variable that indicates whether the object has fraud.	1
	revenue	Amount of tariff raised after the inspection, only available on some illicit cases.	$20

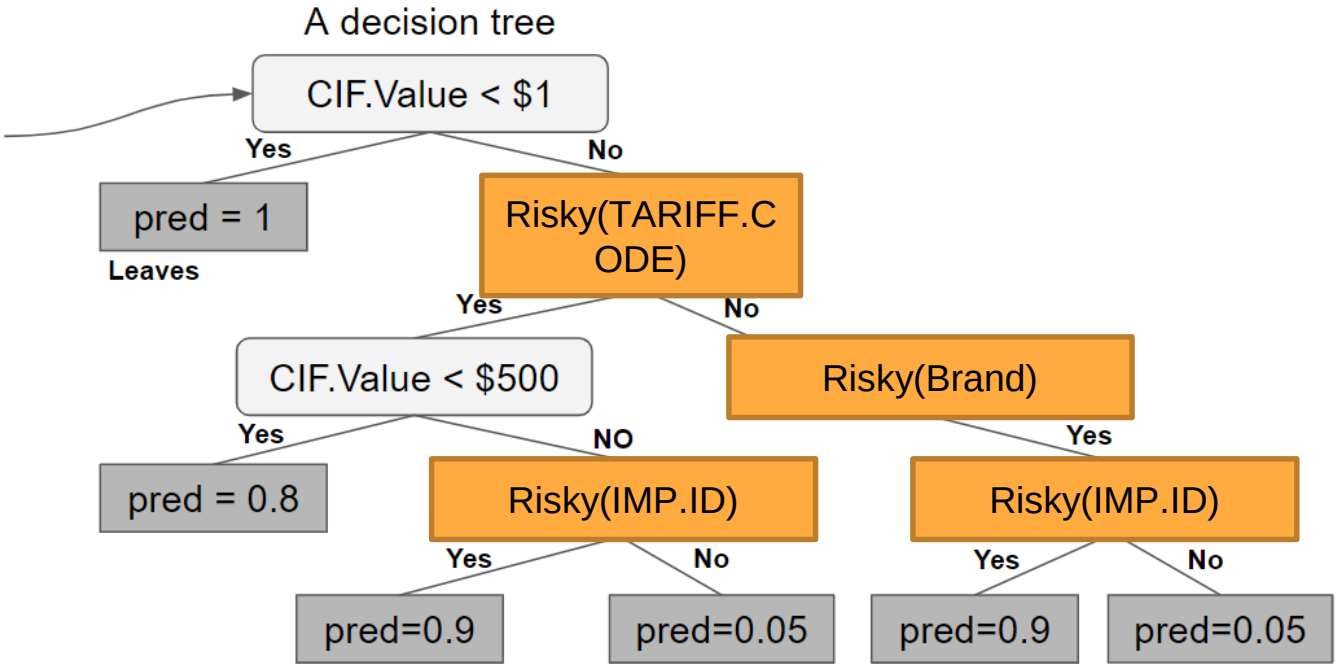
Previous works – Decision trees

Decision criteria
(e.g., HS6)

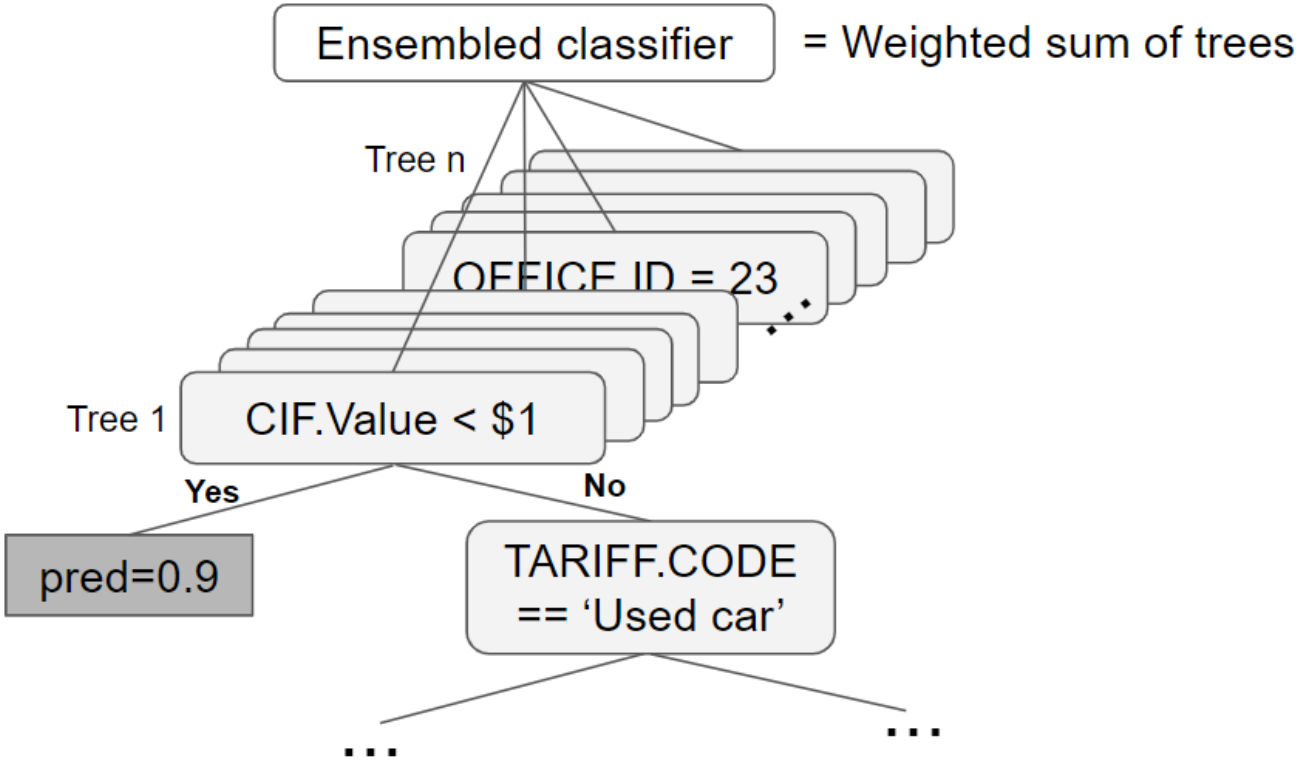


Previous works – Decision trees (Risky profiles)

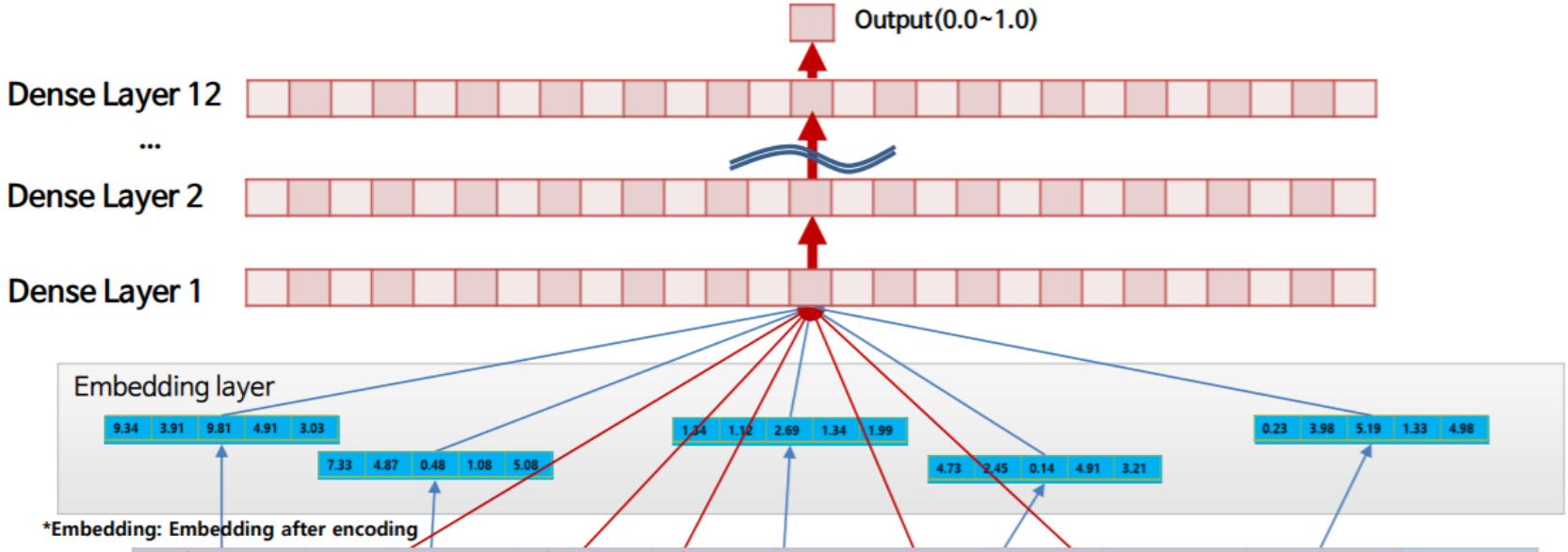
Decision criteria (e.g., HS6)



Previous works – Random forests



Previous works – Fully-connected neural networks



Import declaration number	Column number	Reporter Code	Customs clearance code	Import transaction code	Company Registration Number	...	Reported Product Name	Taxable Price	# of Caught
41436-18-010015M	001	41436	030	11	6059139236	...	OTHER	125,276	1

Image from BACUDA Experts Meeting, 2020

Previous works – Convolutional neural networks

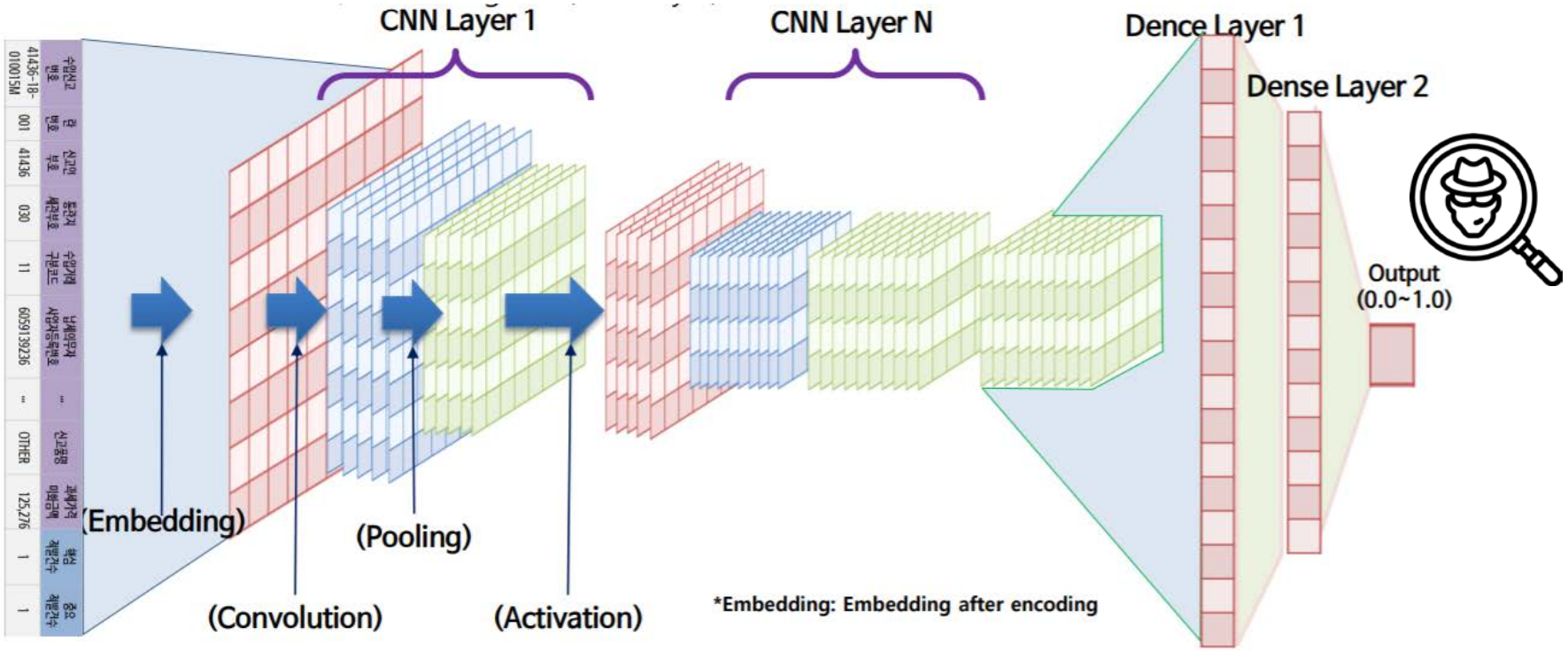
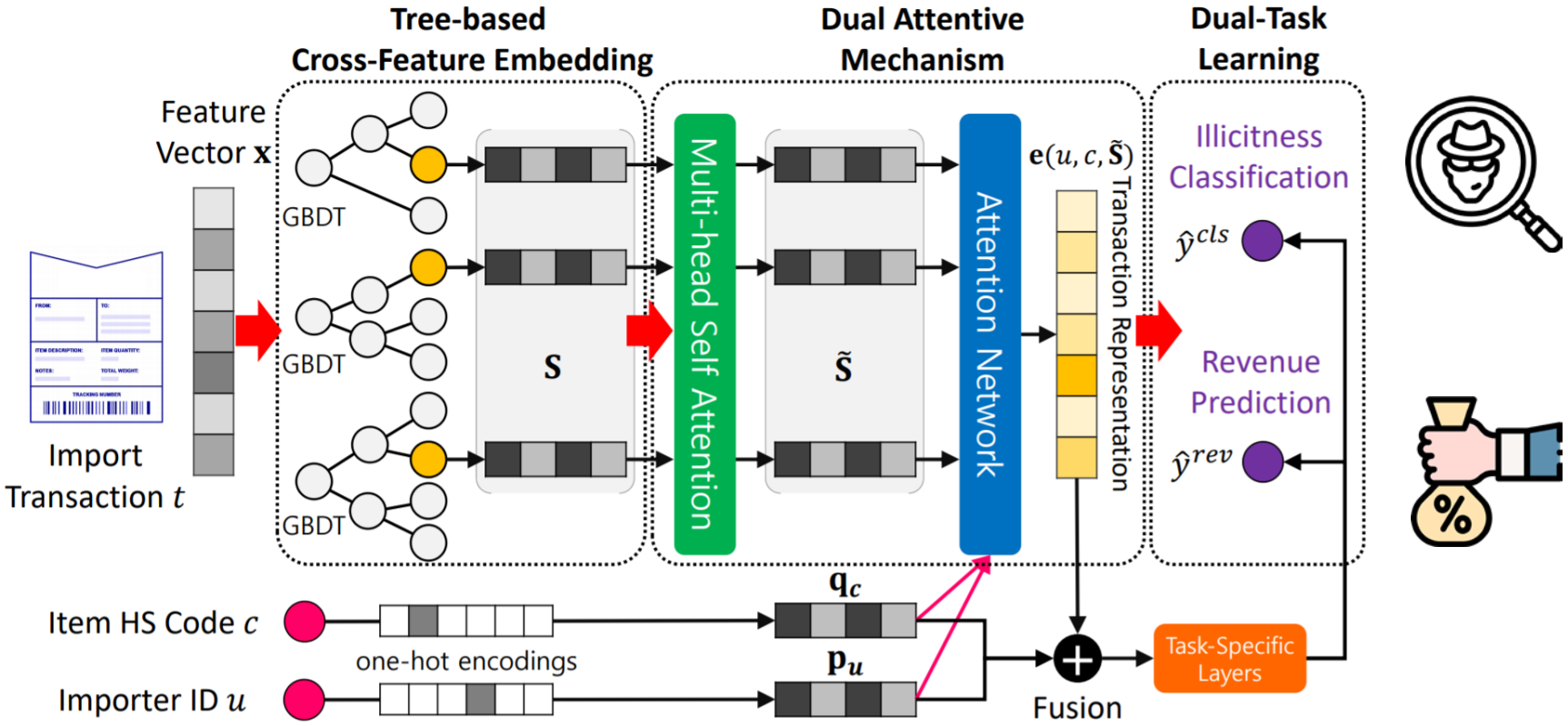


Image from BACUDA Experts Meeting, 2020

Previous works – Dual attentive tree-aware embedding



Published at SIGKDD 2020, http://bit.ly/kdd20-date

Evaluation

- Used data: Nigeria
- Training: Y2013-2016
- Testing: Y2017
- Average Illicit rate: 2.2% (Y2017)

Precision, Recall

Revenue

Model	n = 1% (Selecting top 1%)			n = 2%		
	Pre.	Rec.	Rev.	Pre.	Rec.	Rev.
Price	2.75%	1.23%	15.17%	2.23%	1.99%	20.64%
Importer	11.43%	5.10%	4.36%	9.41%	8.39%	7.56%
IForest	5.61%	2.50%	14.30%	6.19%	5.52%	23.14%
GBDT	90.01%	40.15%	24.59%	66.16%	59.04%	38.89%
GBDT+LR	90.95%	40.40%	27.18%	72.94%	65.09%	44.22%
TEM	88.72%	39.59%	39.48%	74.70%	66.43%	58.48%
DATE$_{CLS}$	**92.66%**	**41.33%**	44.97%	**80.79%**	**72.05%**	67.14%
DATE$_{REV}$	82.25%	36.63%	**49.29%**	79.93%	71.22%	**68.48%**

+

Supporting Experiments

- Ablation analysis
- Effects on training length
- Performance on test subgroups
- Robustness on injected noise

Is there anything else to consider?

Long-term sustainability

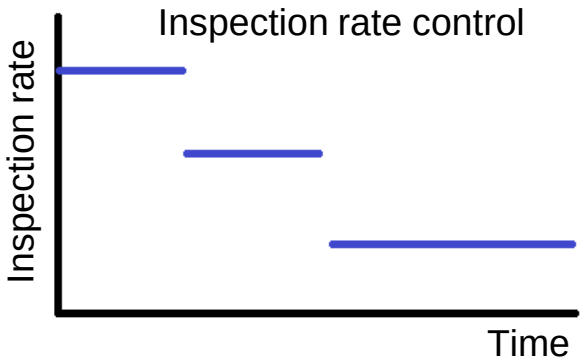
Long-term sustainability

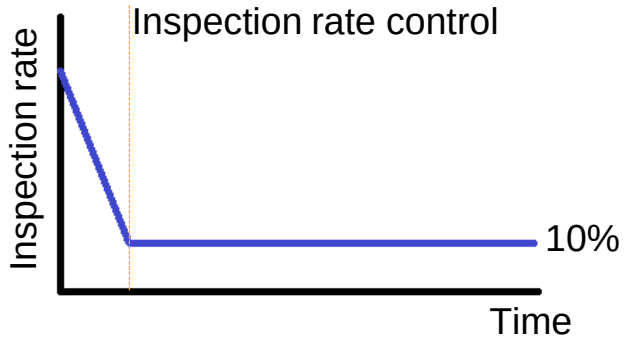
$$f_m = \underset{f}{\operatorname{argmax}} \frac{1}{T - t_0 + 1} \sum_{t=t_0}^{T} m(\mathcal{B}_t^s(f))$$

Find the best selection strategy which can maximize the performance in the last k weeks.
(f: strategy, m:metric, B_t^s: inspected trades)

Mimicking an e-Clearance System

- 1 month training data
- Weekly inspection
- Inspected results are added to the next training stage
- Weekly model update
- 10% target inspection rate
- Allow to mix several strategies





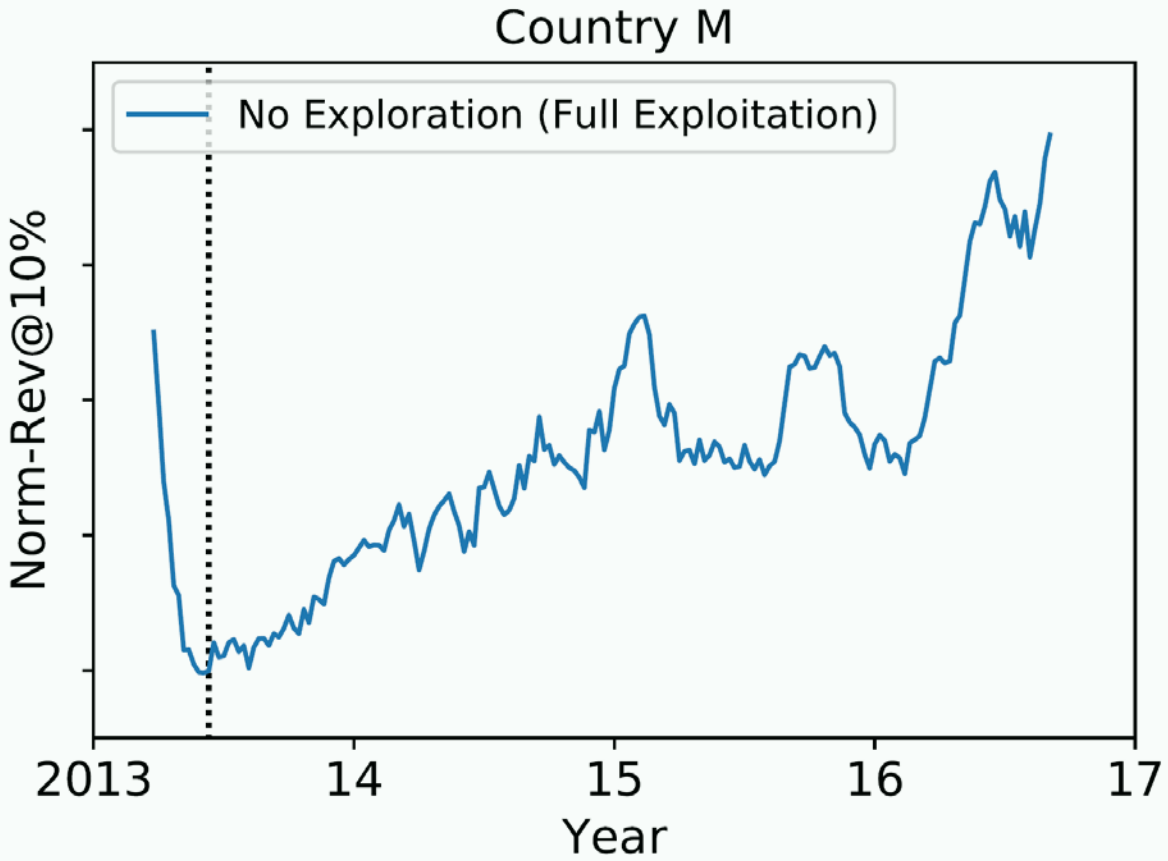
Mimicking an e-Clearance System

- 1 month training data
- Weekly inspection
- Inspected results are added to the next training stage
- Weekly model update
- 10% target inspection rate
- Allow to mix several strategies

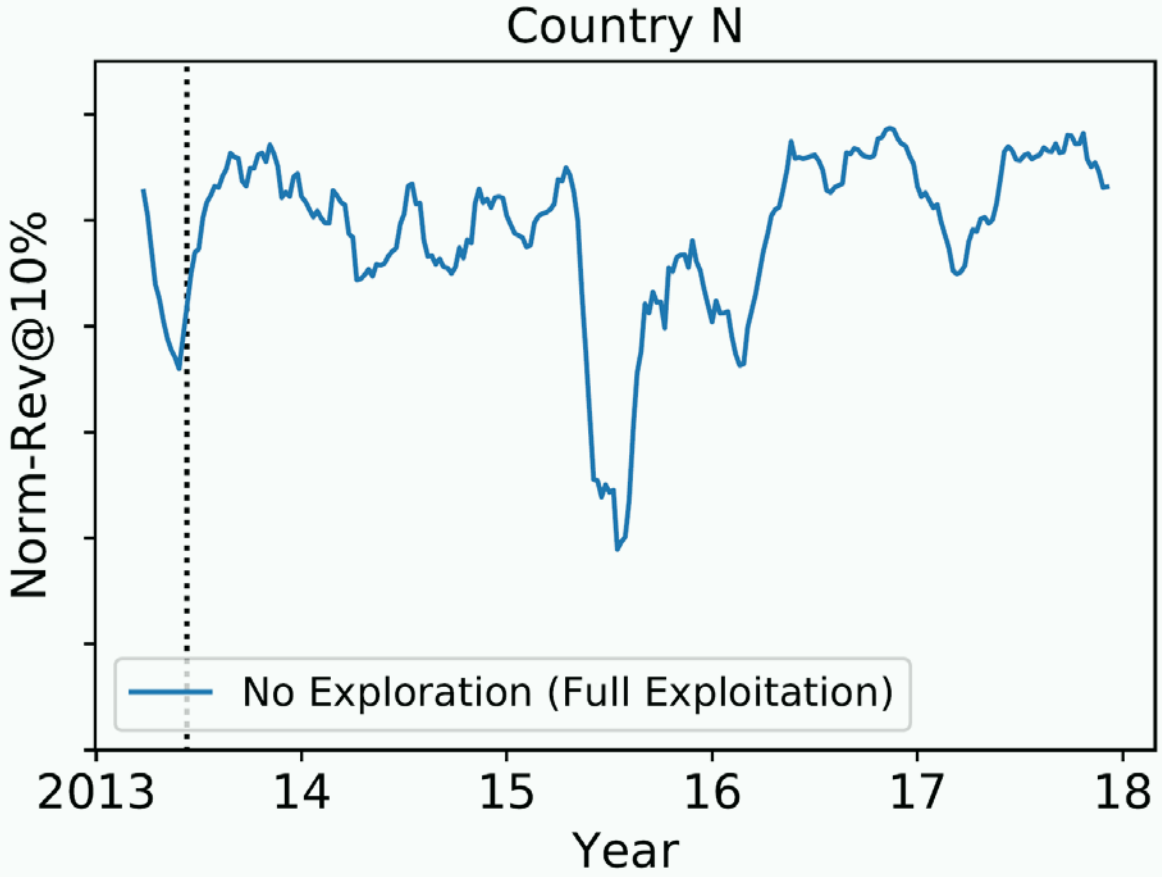
Everything is tunable 

```
$ export CUDA_VISIBLE_DEVICES=3 && python main.py --data real-k --
semi_supervised 0 --sampling hybrid --subsamplings bATE/DATE --weights
0.1/0.9 --mode scratch --train_from 20130101 --test_from 20130701 --
test_length 30 --valid_length 30 --initial_inspection_rate 20 --
final_inspection_rate 5 --epoch 5 --closs bce --rloss full --save 0 --
numweeks 100 --inspection_plan direct_decay
```

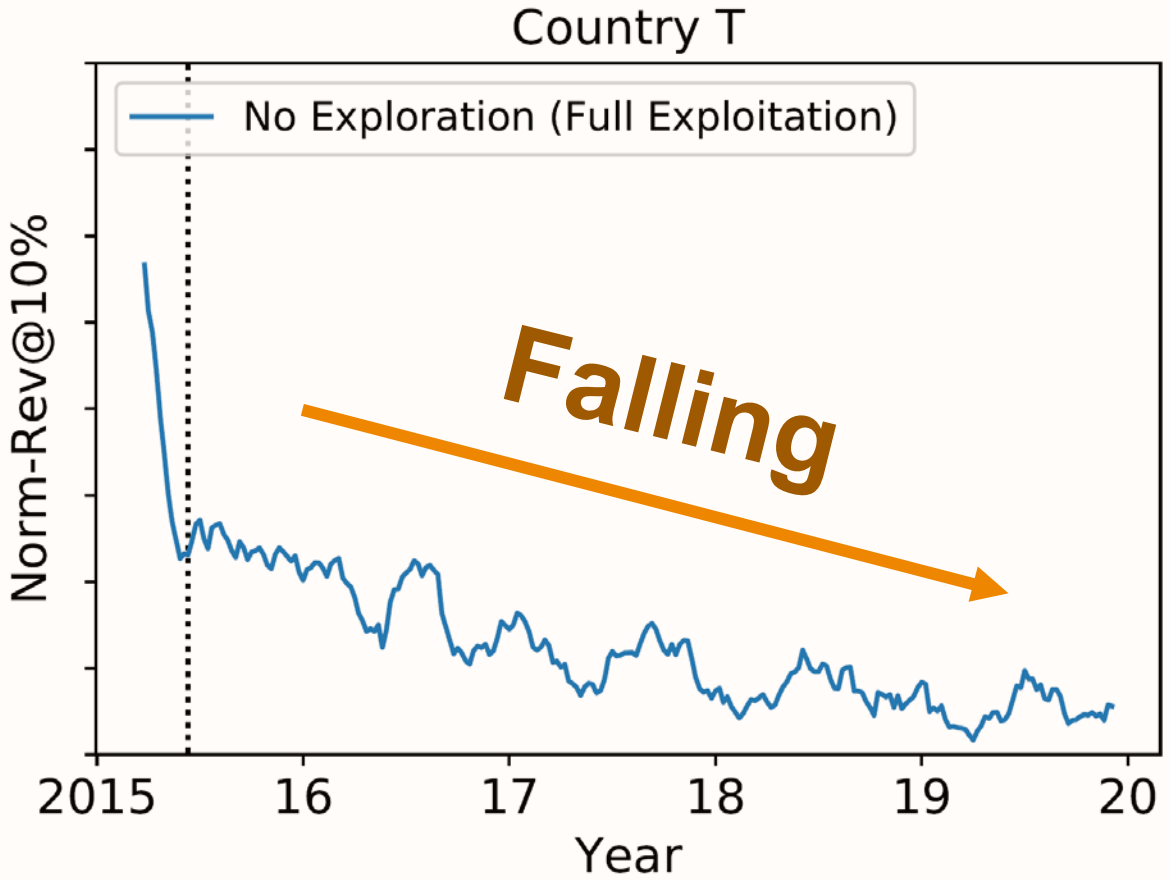
Sustainability of the Algorithm



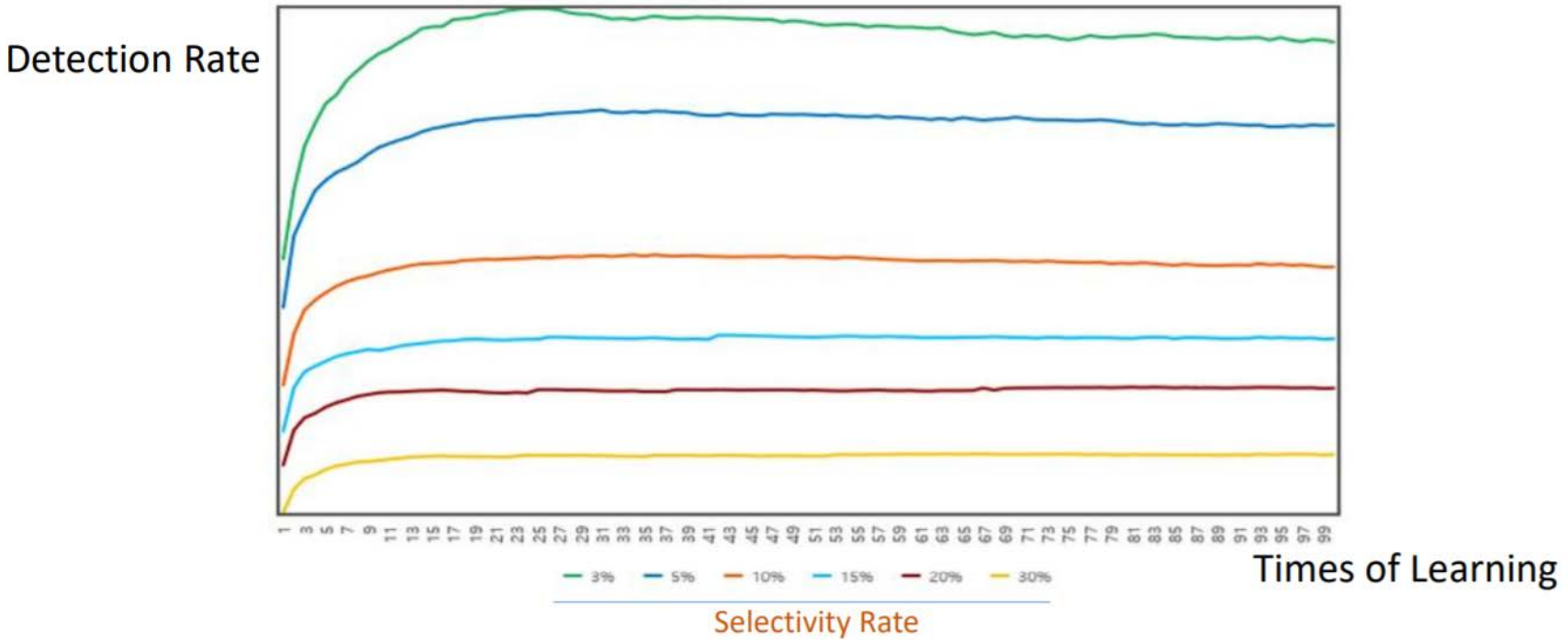
Sustainability of the Algorithm



Sustainability of the Algorithm



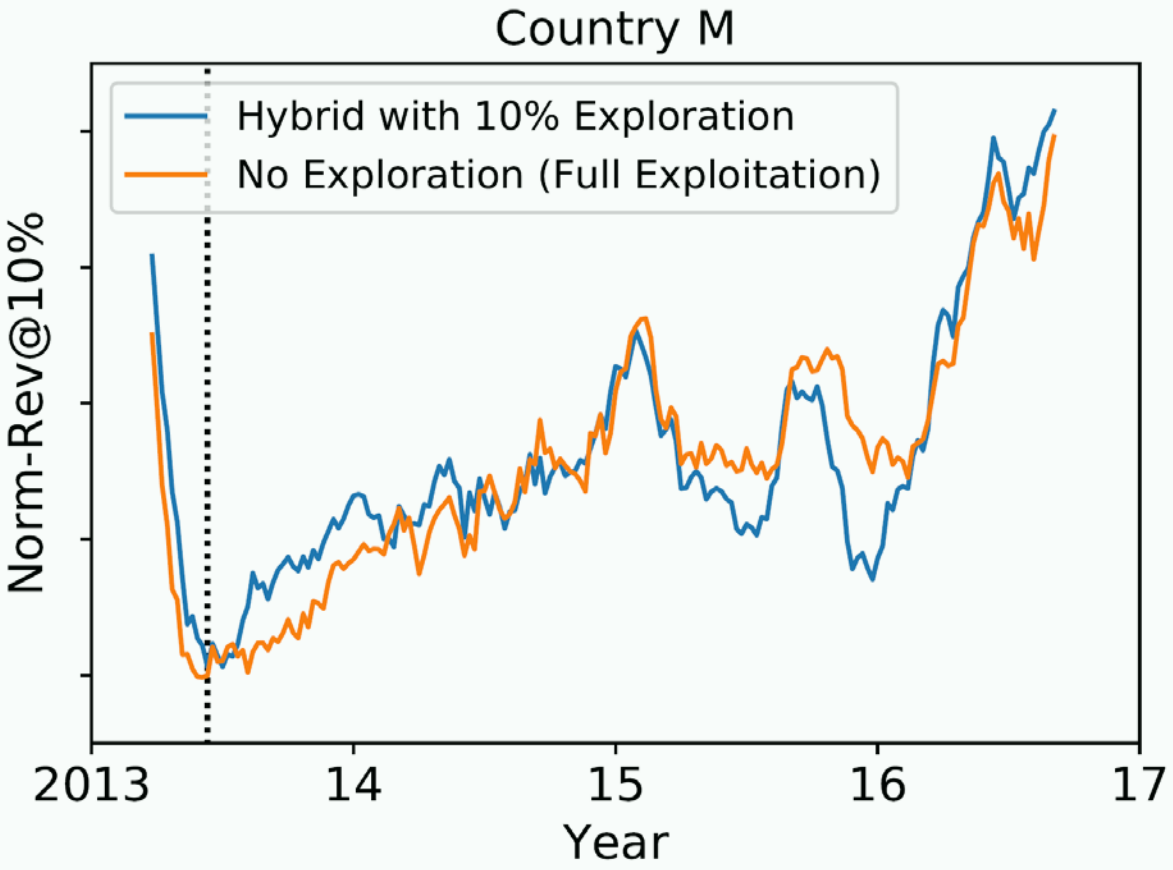
Sustainability of the Algorithm



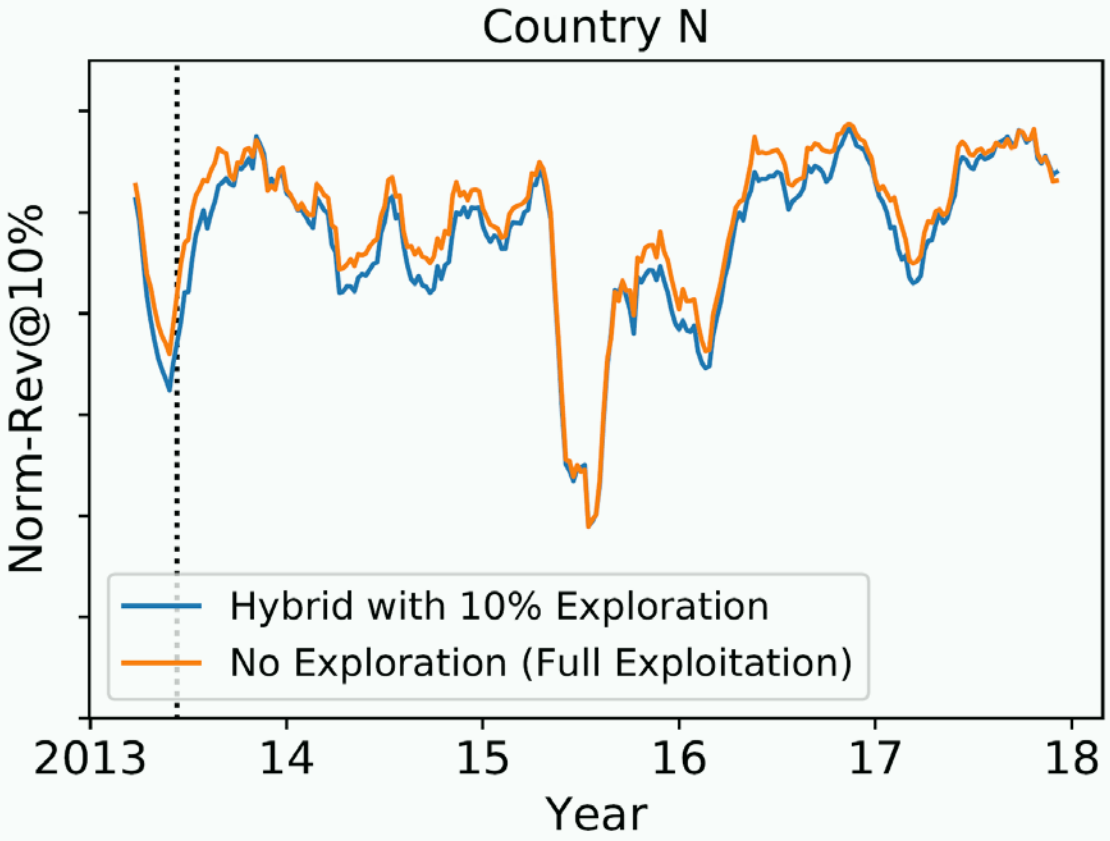
Excerpted from BACUDA Experts Meeting, 2020

What would be the reason? How can we solve?

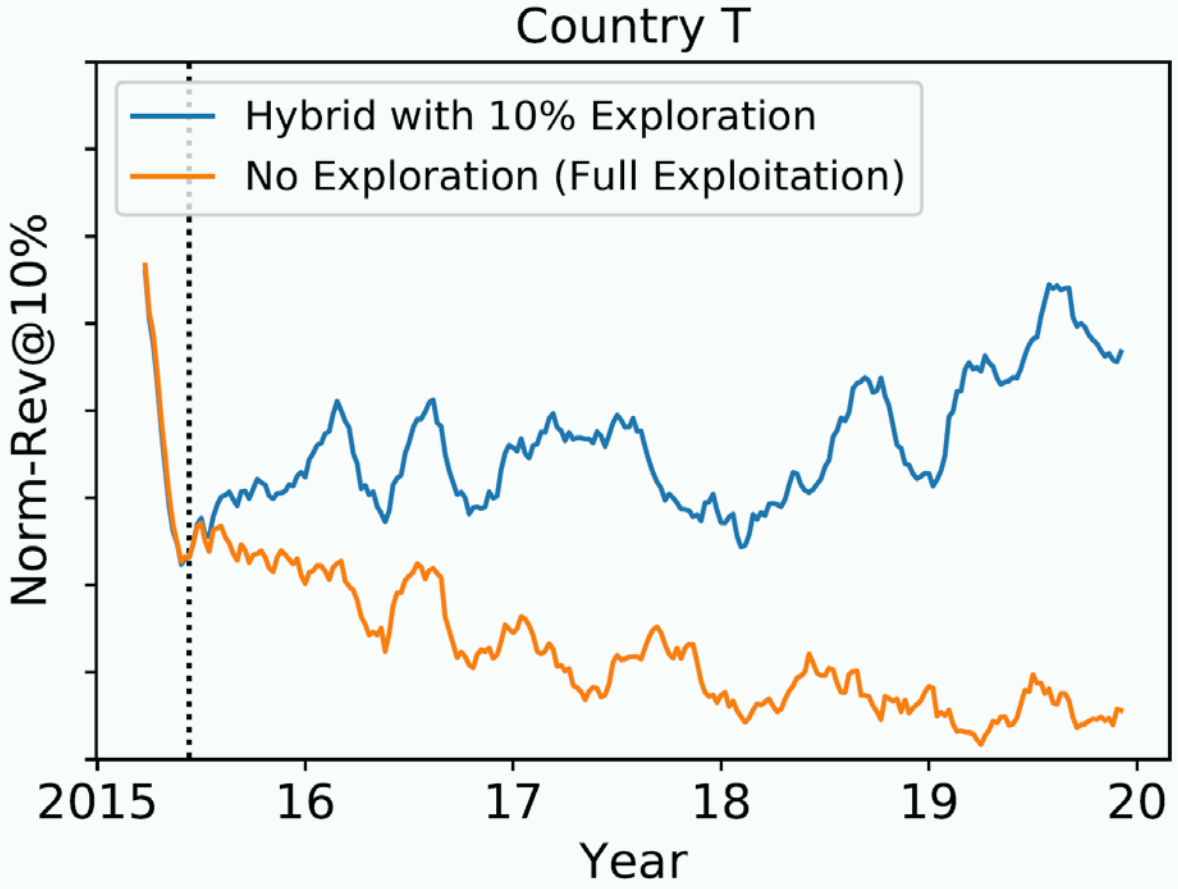
Do we need some exploration?



Do we need some exploration?



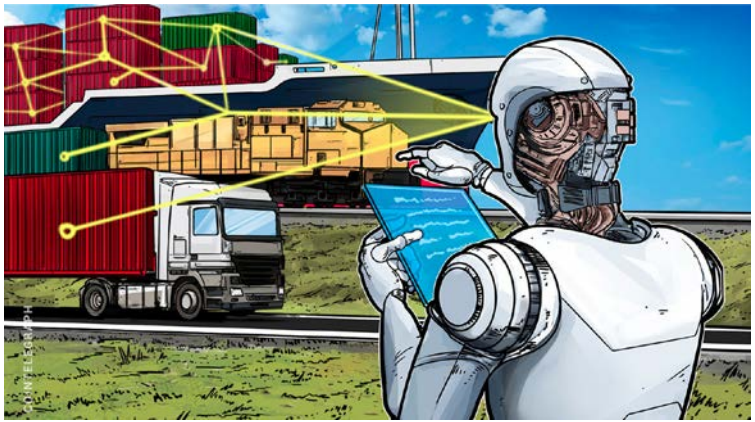
Some exploration helps!



What would be the reason?

Active Learning

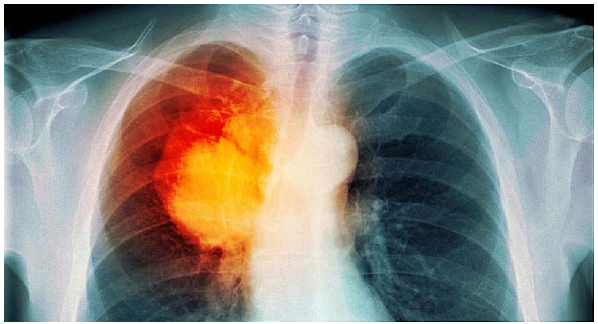
- Active learning is a branch of machine learning to find the most efficient data sample to train a model
- It is used widely when annotating data samples is expensive especially if each annotation task relies on subject matter expert’s knowledge.



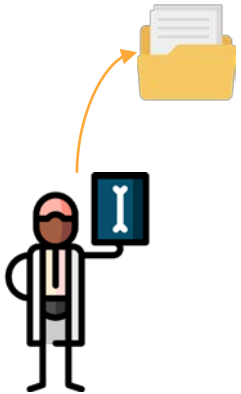
AI will support labor-intensive tasks, but training an AI requires lots of efforts.

Active Learning

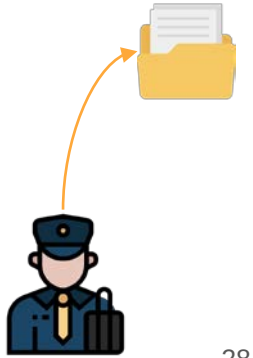
- Active learning is a branch of machine learning to find the most efficient data sample to train a model
- It is used widely when annotating data samples is expensive especially if each annotation task relies on subject matter expert's knowledge.



Annotating tumors

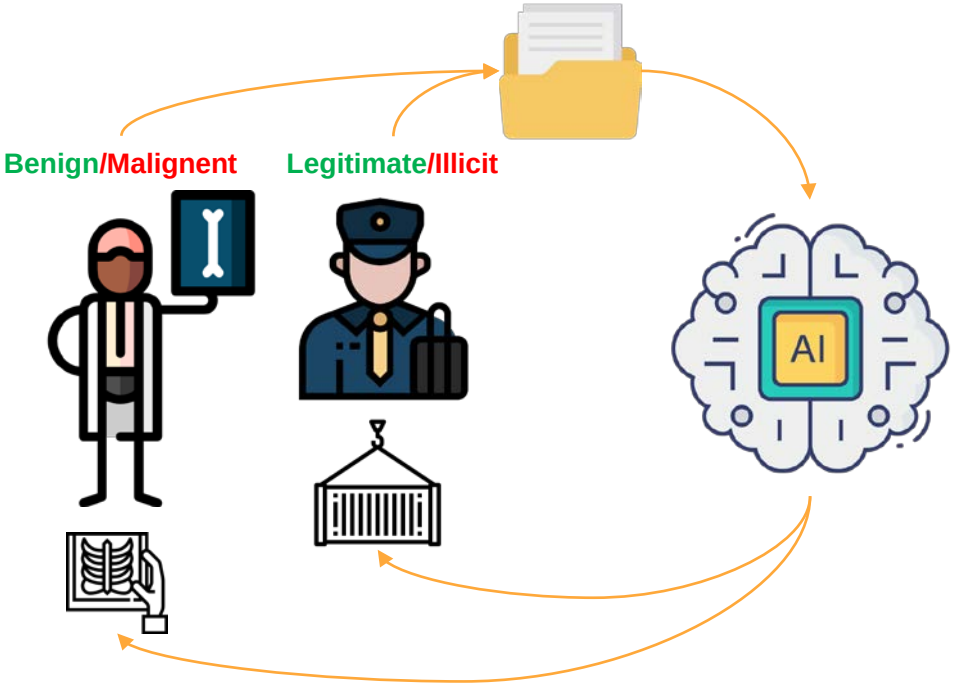


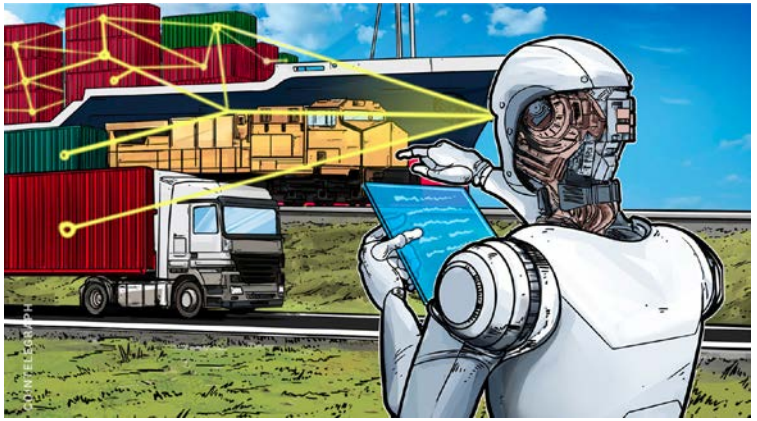
Inspecting frauds



Decision

- To innovate AI, which items should we ask experts to annotate?





http://english.customs.gov.cn/statics/6483b197-5ea9-4f90-a04a-89bc9e28de74.html

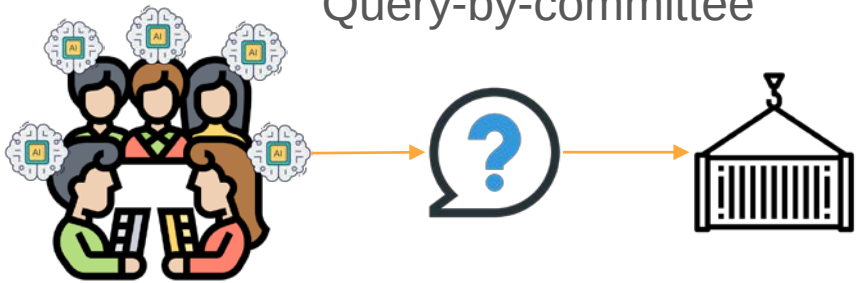
Selection criteria for active learning

- To train the model efficient and effective, diverse approaches are proposed.

Uncertainty-based

Using decision boundary

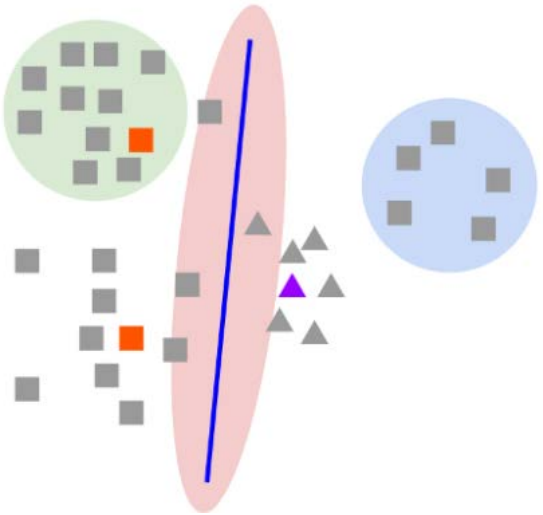
Query-by-committee



Selection criteria for active learning

- To train the model efficient and effective, researchers studied diverse approaches.

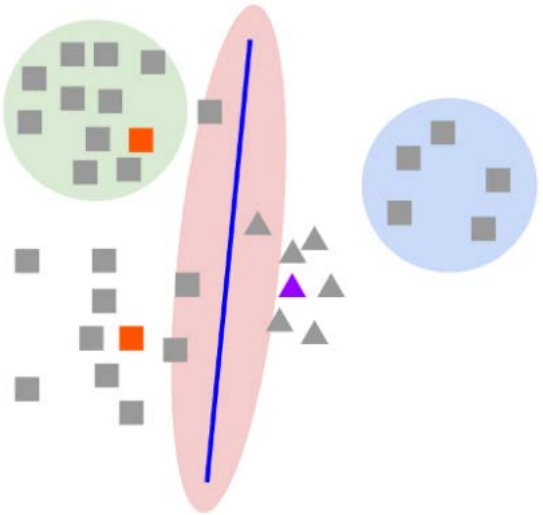
Diversity-based



Pool-based active learning

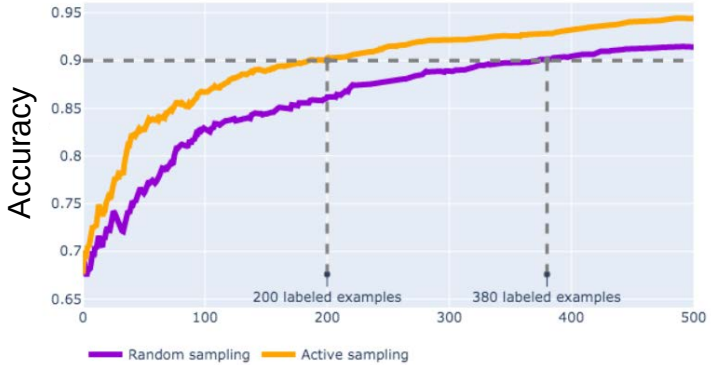
- To train the model efficient and effective, researchers studied diverse approaches.

Diversity-based

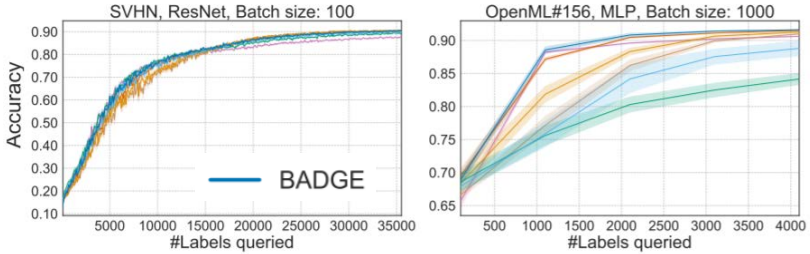


Effect of active learning (On benchmark datasets)

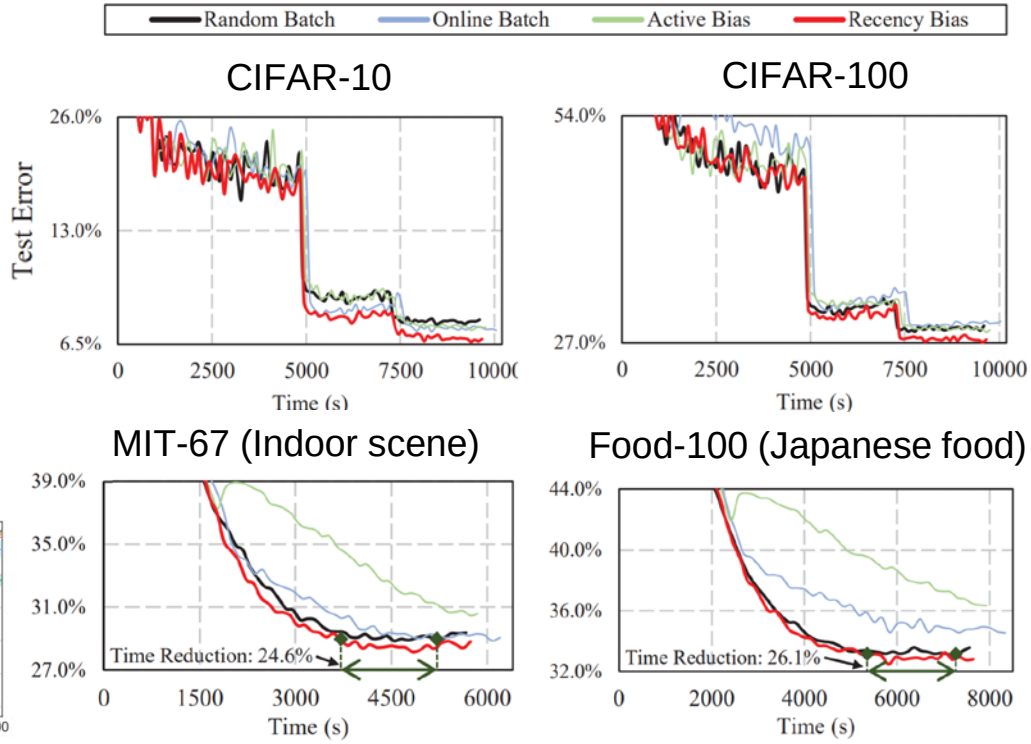
Fruit fresh-rotten classification



https://www.kaggle.com/sriramr/fruits-fresh-and-rotten-for-classification



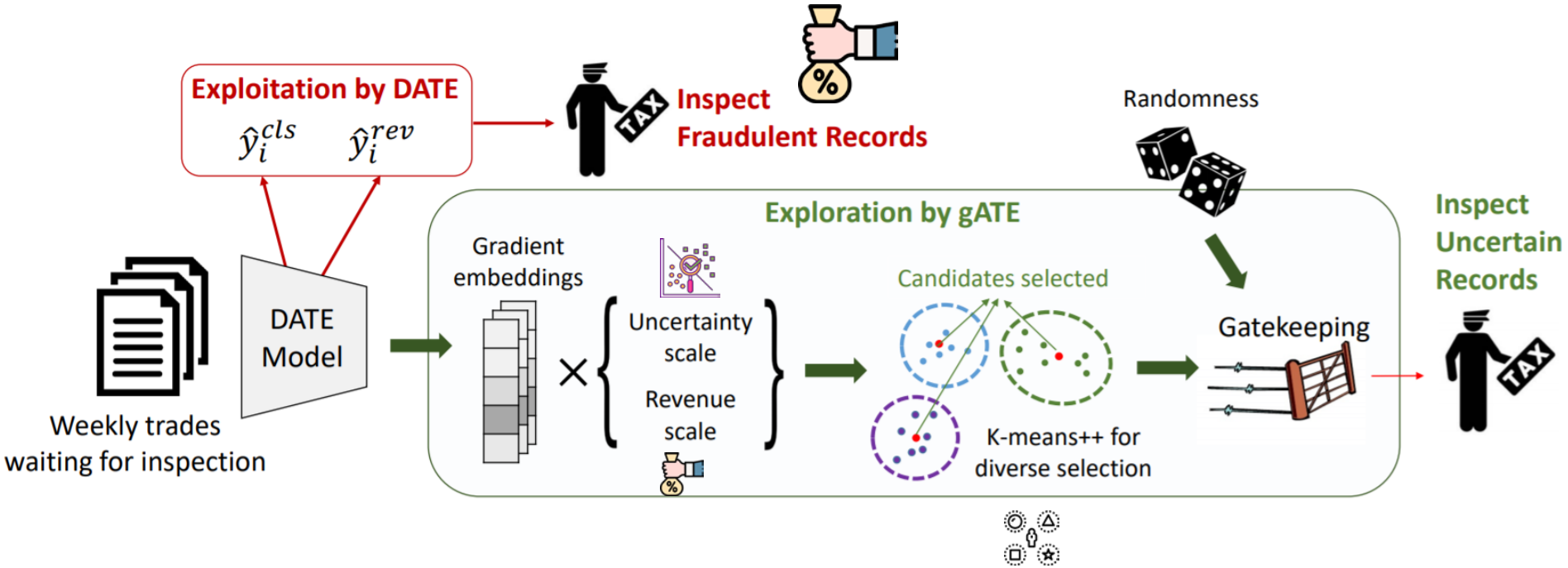
BADGE, ICLR 2020 https://arxiv.org/pdf/1906.03671.pdf



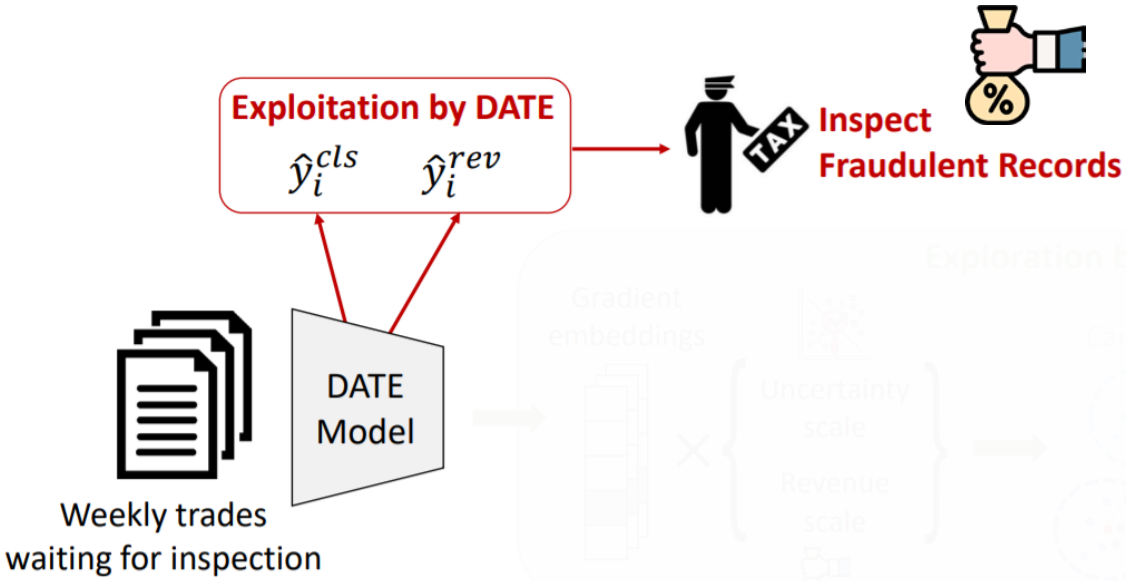
Recency Bias, CIKM 2020
https://dl.acm.org/doi/abs/10.1145/3340531.3411898

Come back to our problem setting

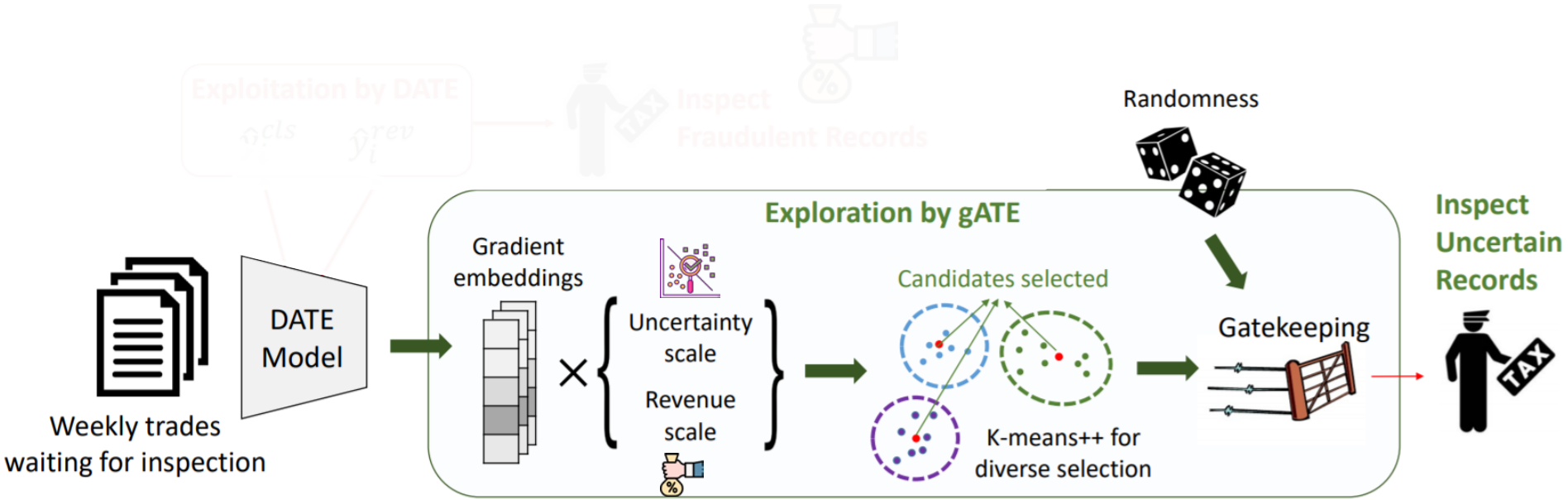
Advancing customs selection with active learning



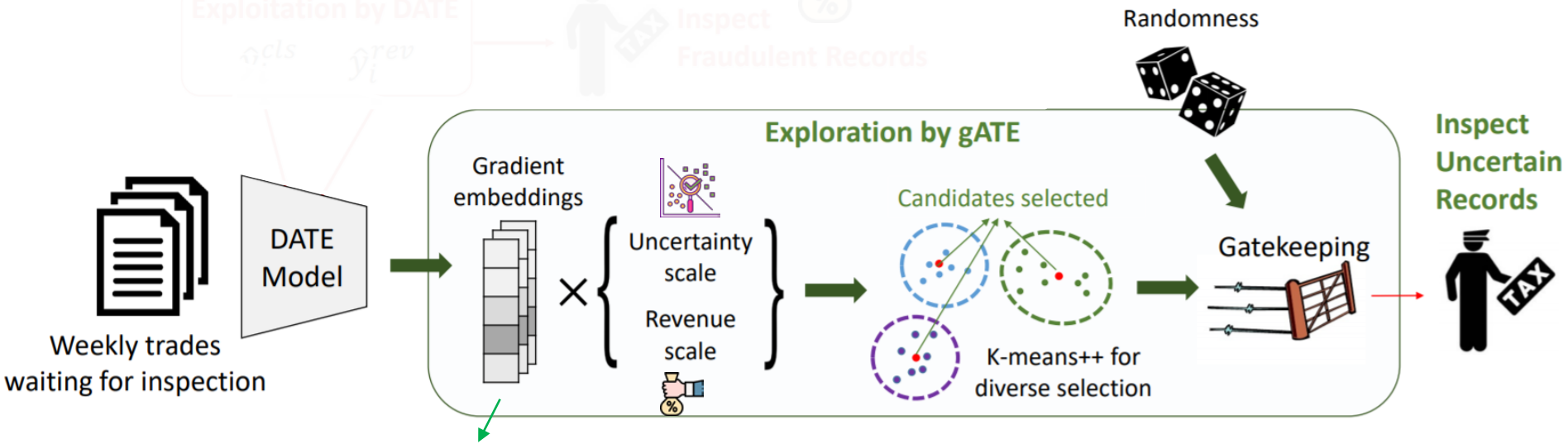
Selecting fraudulent items by exploitation



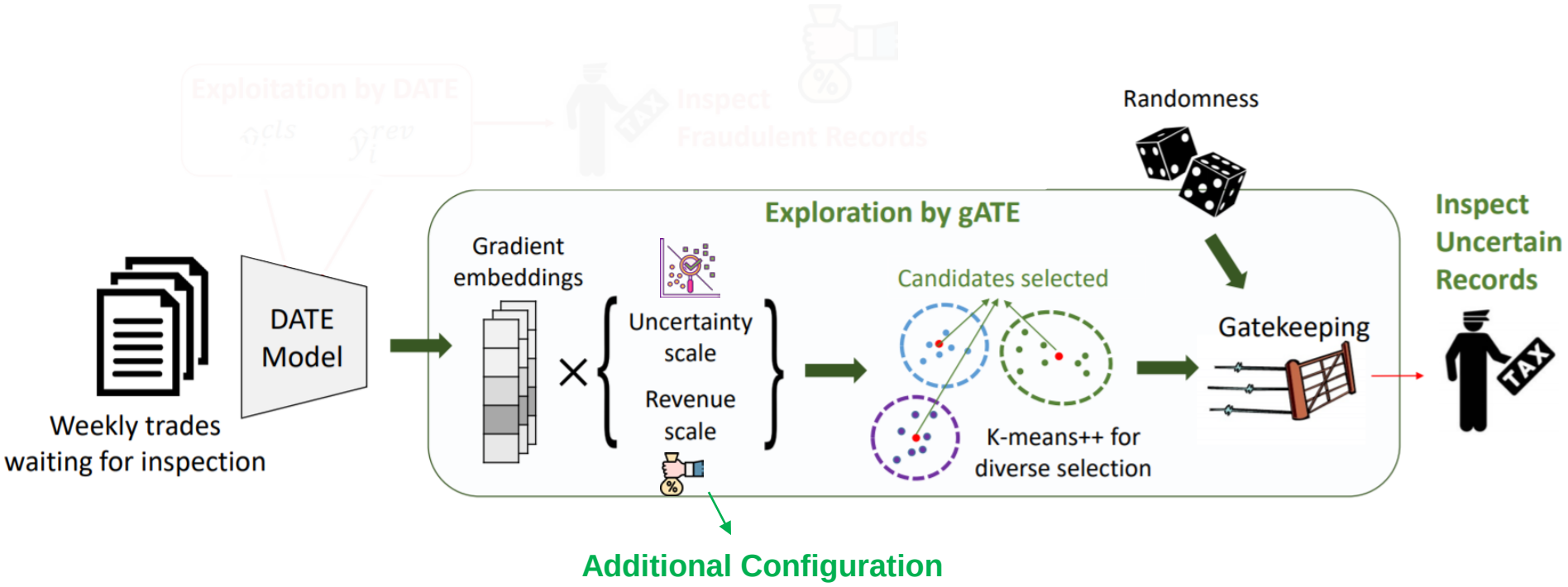
Selecting uncertain items by exploration (gATE)



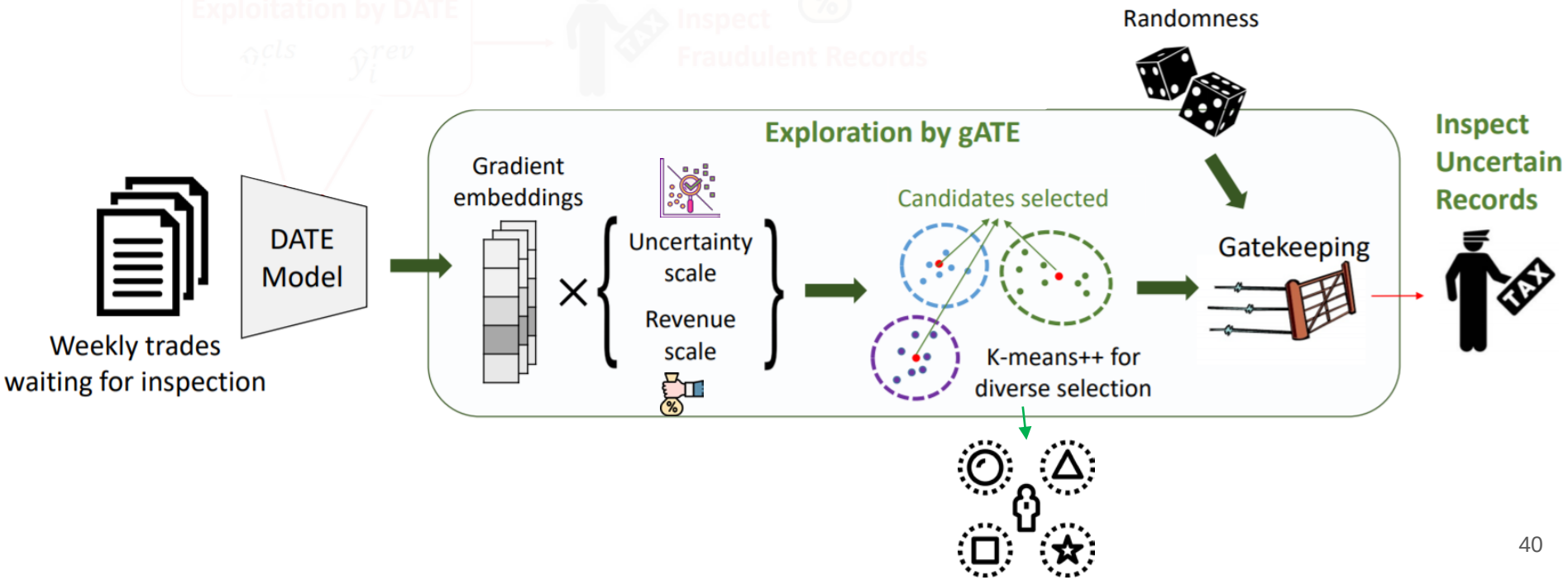
gATE components – Gradient embedding



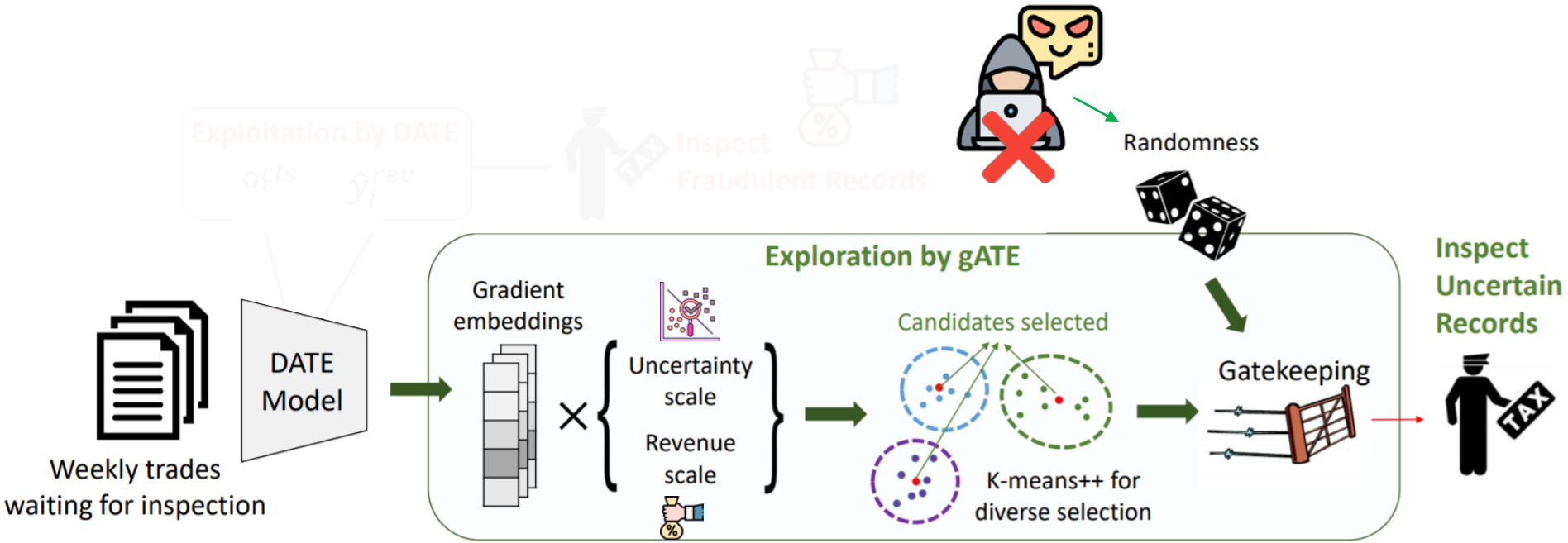
gATE components – Scaling



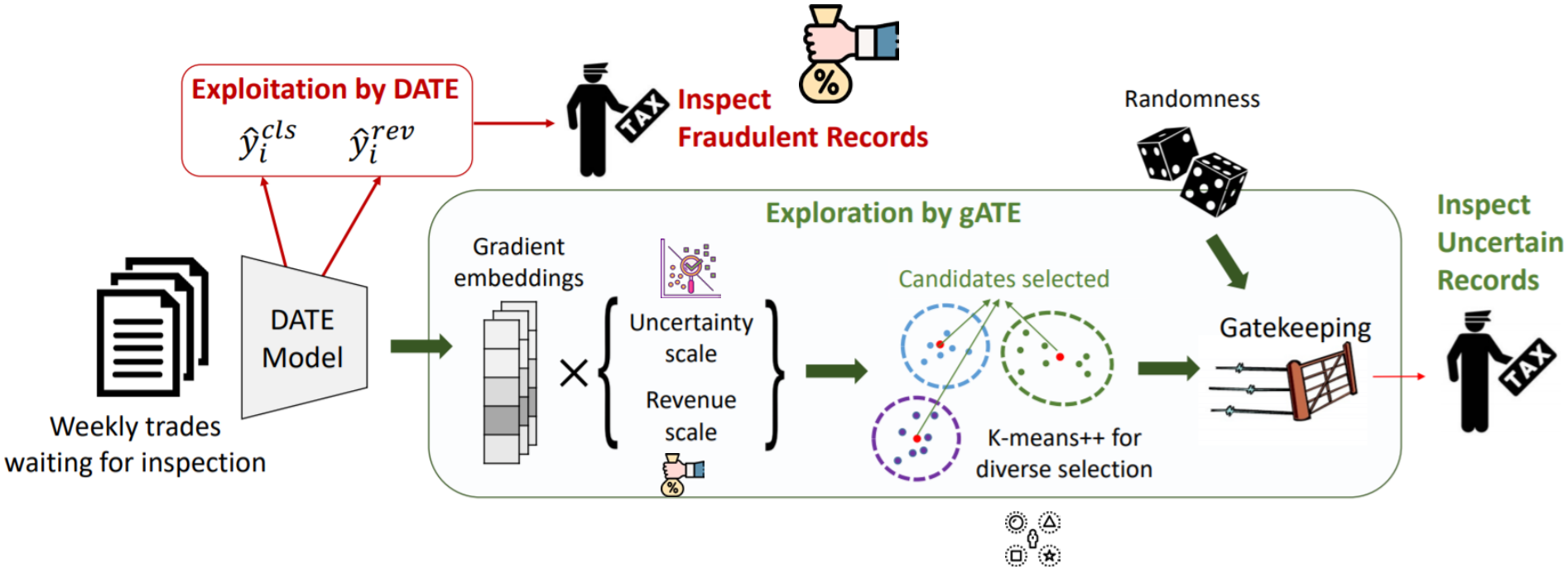
gATE components – Diversifying



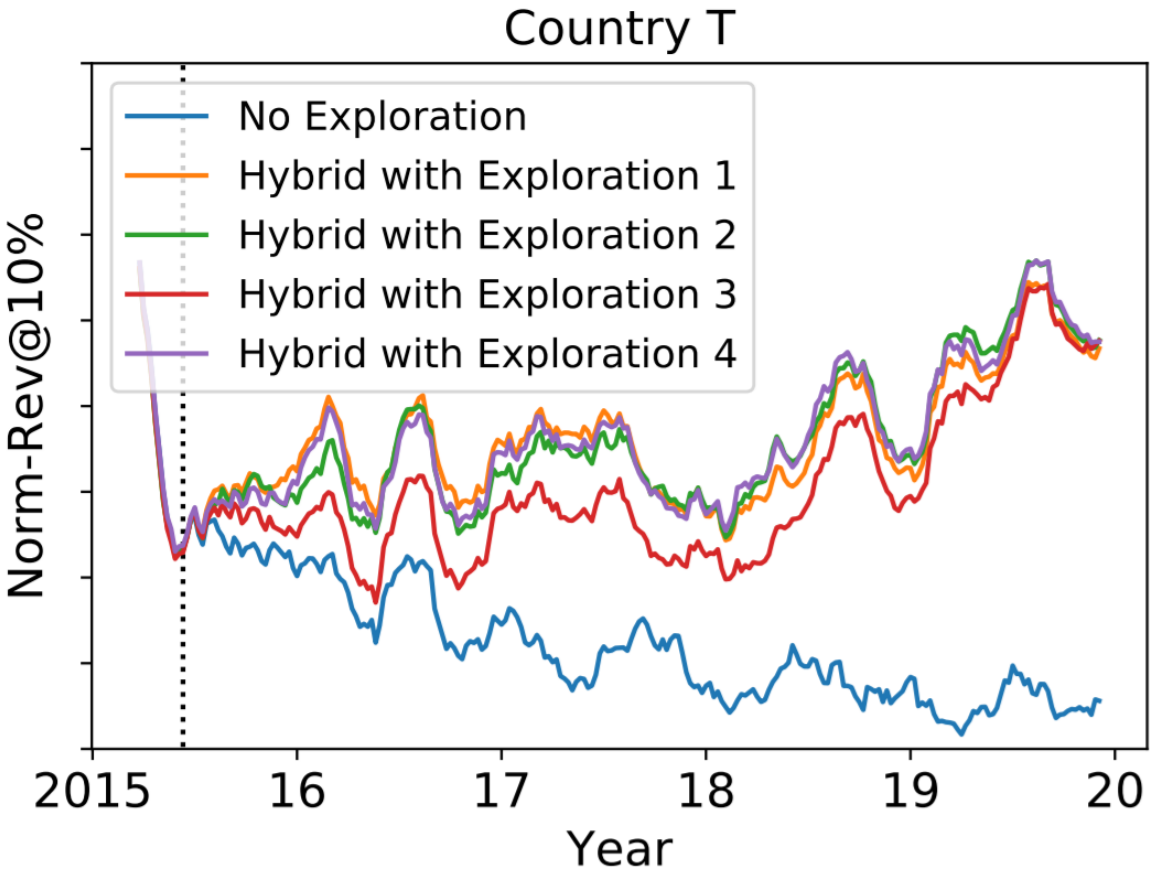
gATE components – Protect adversaries



Proposed hybrid approach



Explored diverse exploration strategies

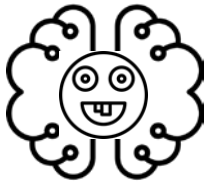


Explored diverse exploration strategies

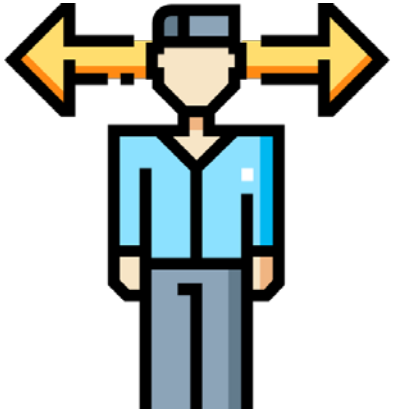
Model	Country M	Country N	Country T
No Exploration	-6.4%	-0.1%	**-74.8%**
Random	0%	0%	0%
BADGE (ICLR'20)	-0.9%	0.8%	2.2%
bATE (Ours)	**3.6%**	0.7%	0.0%
gATE (Ours)	-1.0%	0.7%	**2.9%**

Exploitation-exploration dilemma

Short-term Revenue (Exploitation)



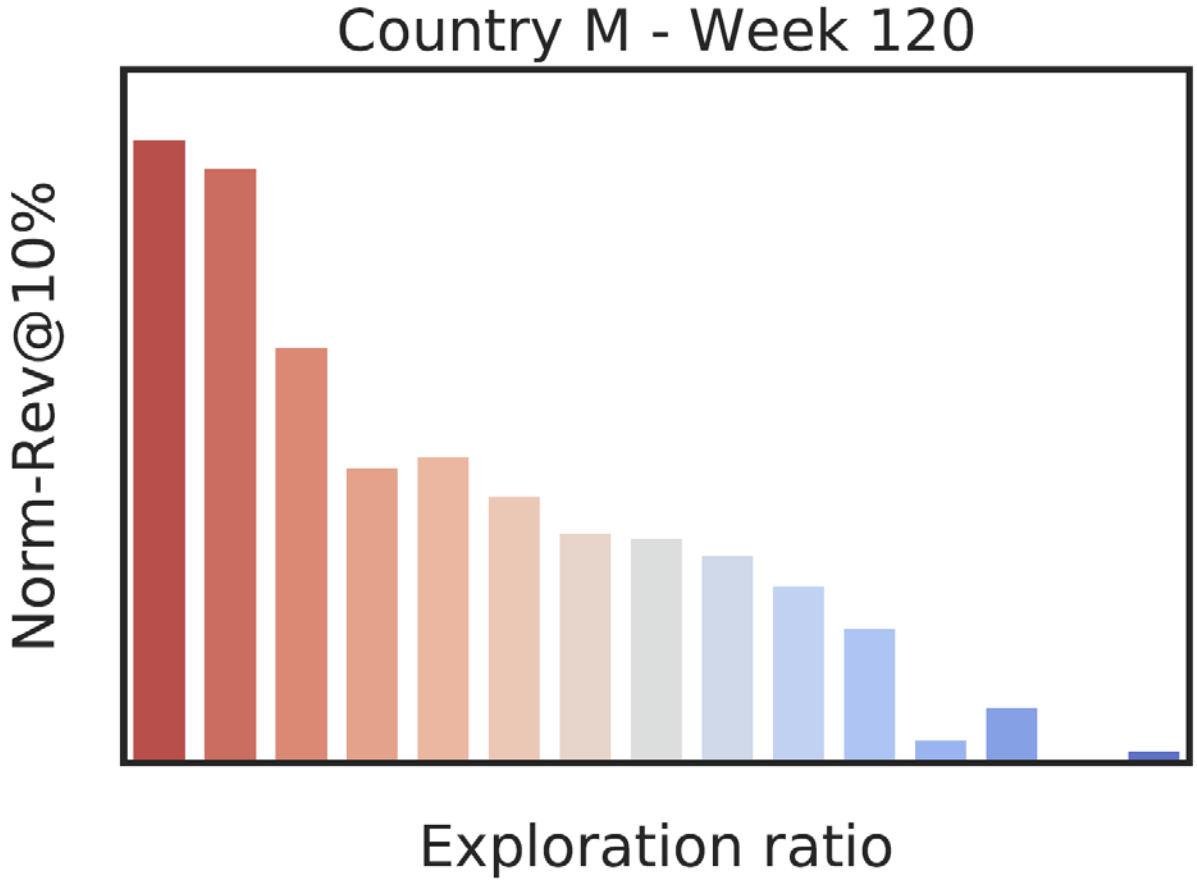
Customs Expert



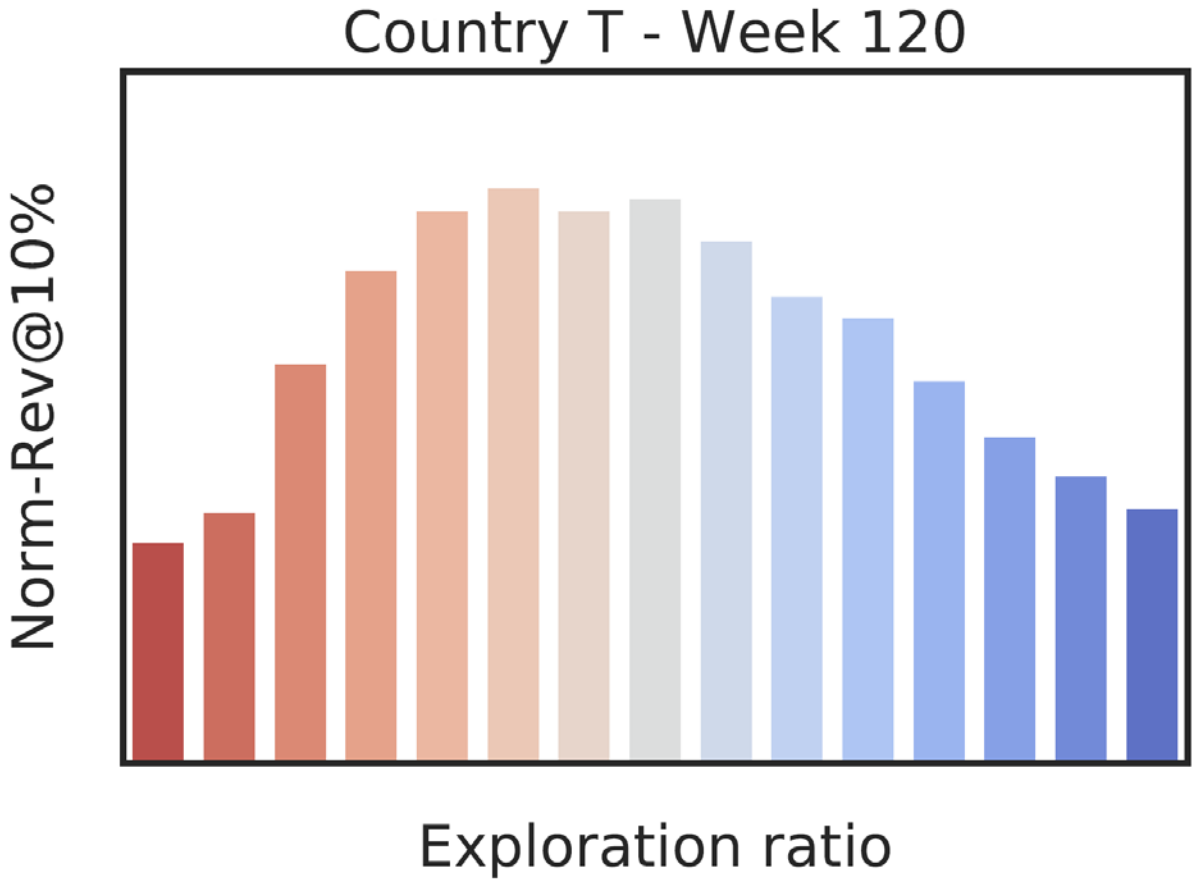
Long-term Advancement (Exploration)

How can we mitigate the tradeoff?

Finding the best exploration ratio



Finding the best exploration ratio



Advancing AI by using unlabeled data together

Uninspected: 95%

Inspected: 5%

+

Hierarchical Classification

Definition & Example for U.S. HTS Codes

[hts code example]

0901.21.0010

What these numbers mean

09 — *Coffee, Tea, Mate And Spices*
Chapter

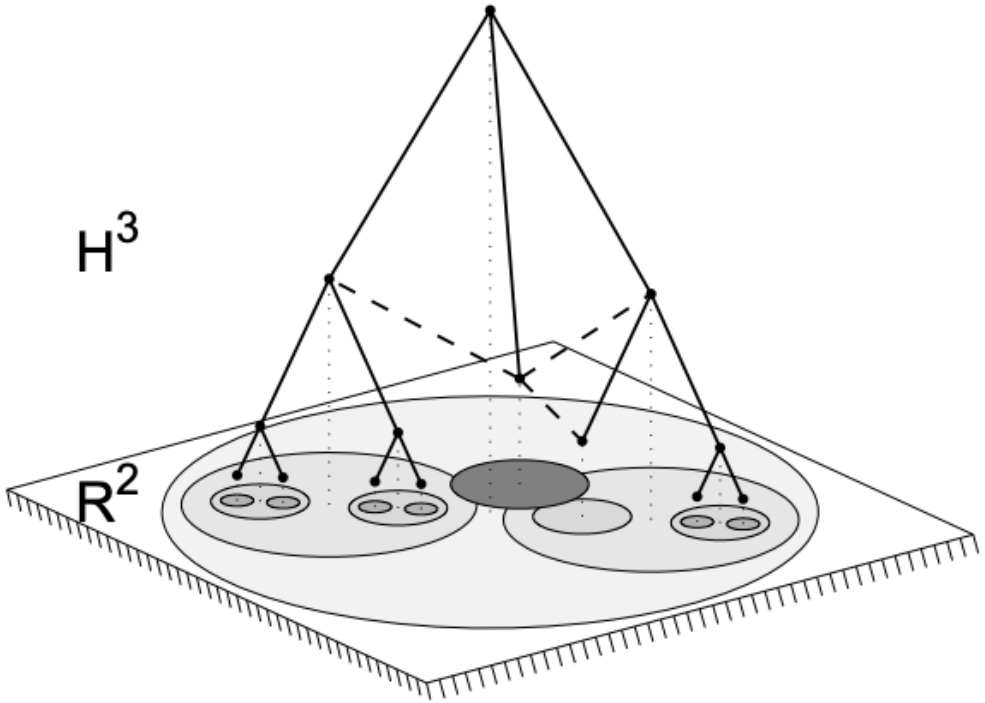
0901 — *Coffee, Whether Or Not Roasted Or Decaffeinated; Coffee Husks And Skins; Coffee Substitutes Containing Coffee*
Heading

0901.21 — *Coffee, Roasted, Not Decaffeinated*
Sub Heading
(HS code)

0901.21.00 — *No Distinction*
Subheading
(Determines Duty)

0901.21.0010 — *Coffee, Roasted, Not Decaffeinated, Certified Organic*
Statistical Suffix
(Further Definition and Makeup)

DESCARTES Datamyne



Thank you

sundong@ibs.re.kr