Try to Poison My Deep Learning Data? Nowhere to Hide Your Trajectory Spectrum!

Yansong Gao[1,3], Huaibing Peng[2], Hua Ma[3], Zhi Zhang[1], Shuo Wang[4], Rayne Holland[3], Anmin Fu[2], Minhui Xue[3], Derek Abbott[5]

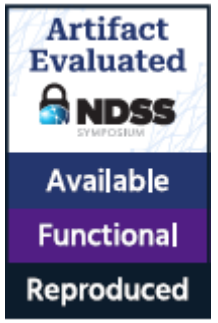
[1]The University of Western Australia

[2]Nanjing University of Science and Technology

[3]CSIRO's Data61

[4]Shanghai Jiao Tong University

[5]The University of Adelaide

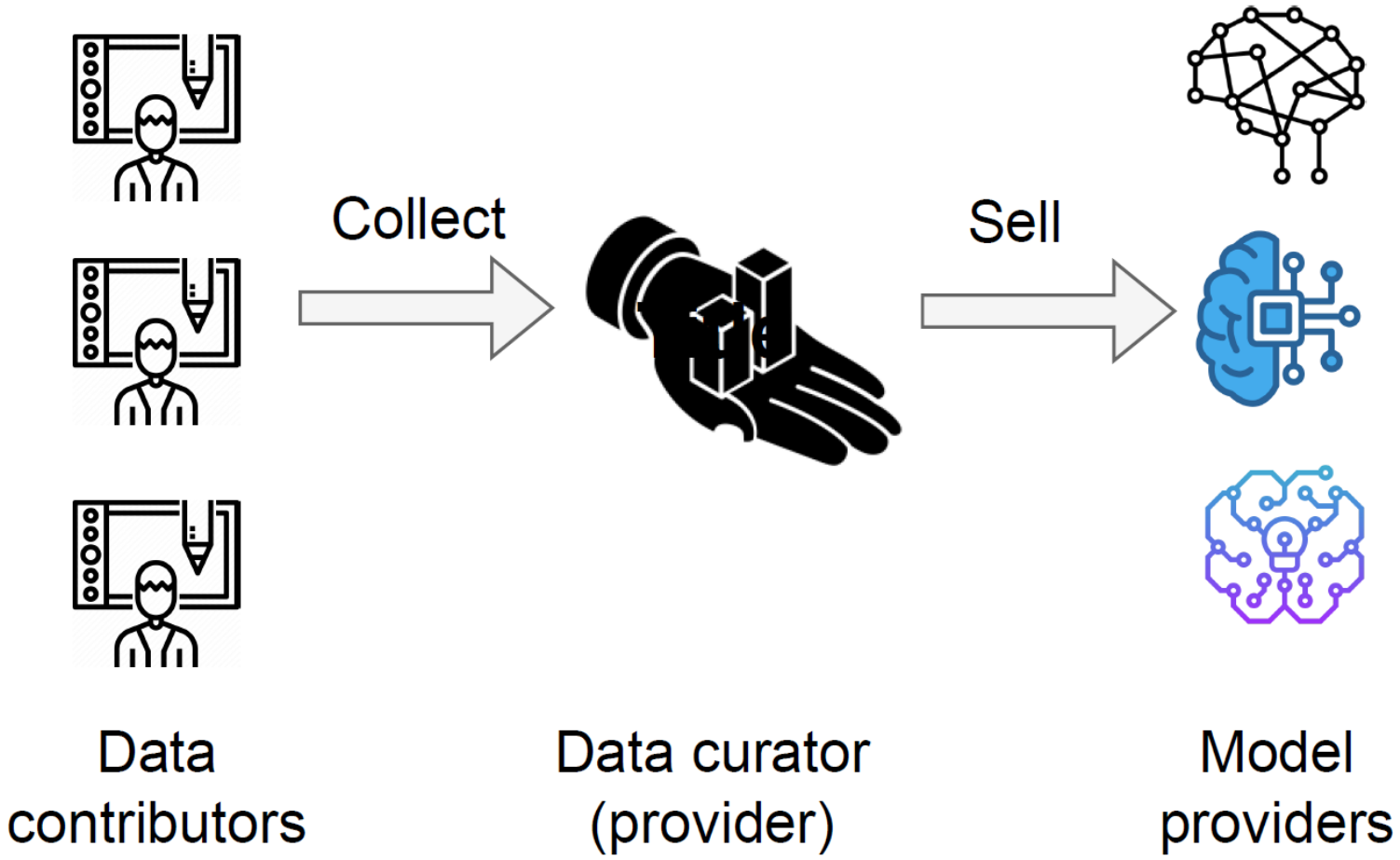


Data as a Service

High-quality data is important!

Acquisition is however challenging

Data as a Service



Data as a Service

appen

e.g., 6 million freelancers

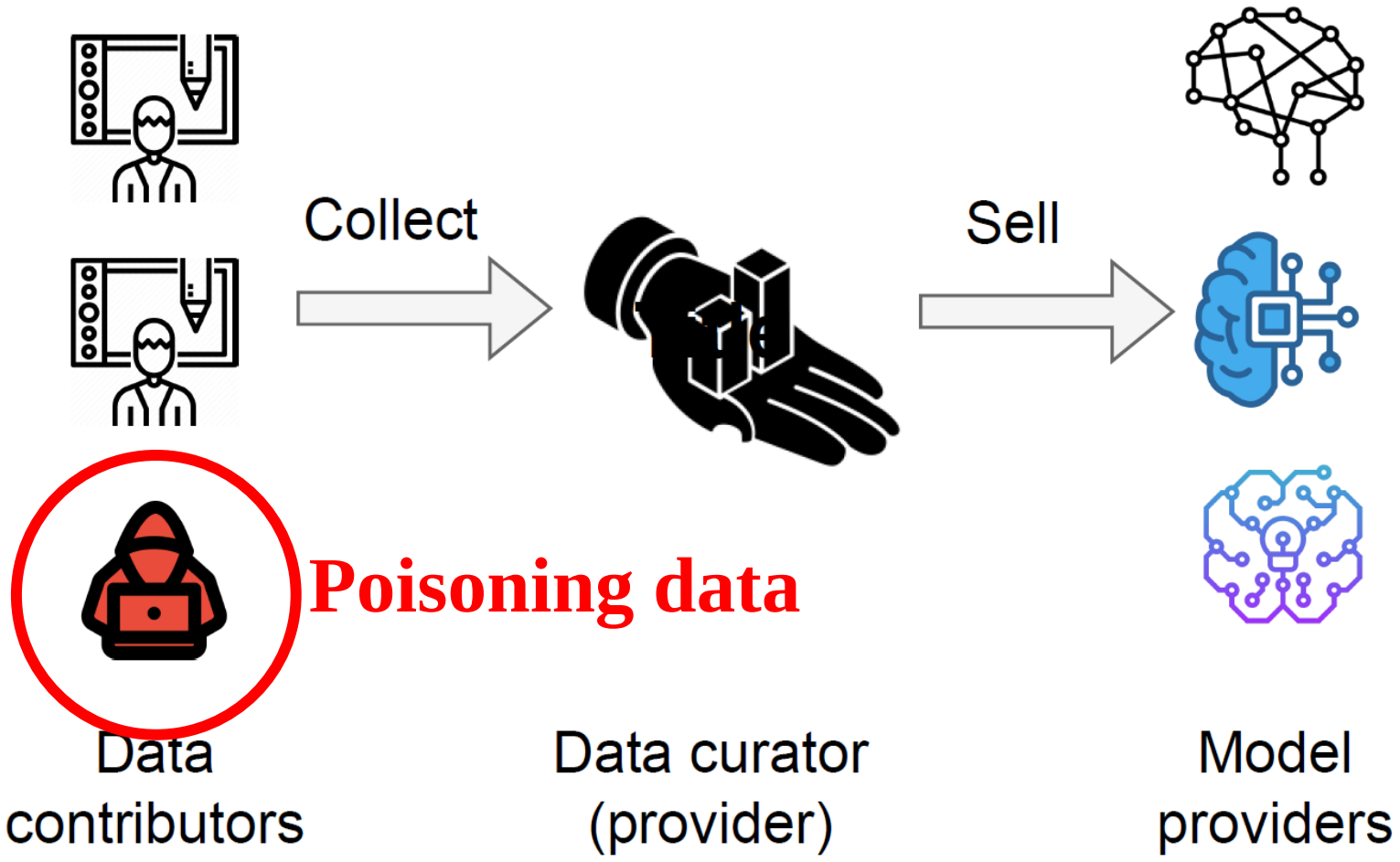
clickworker

amazon mechanical turk

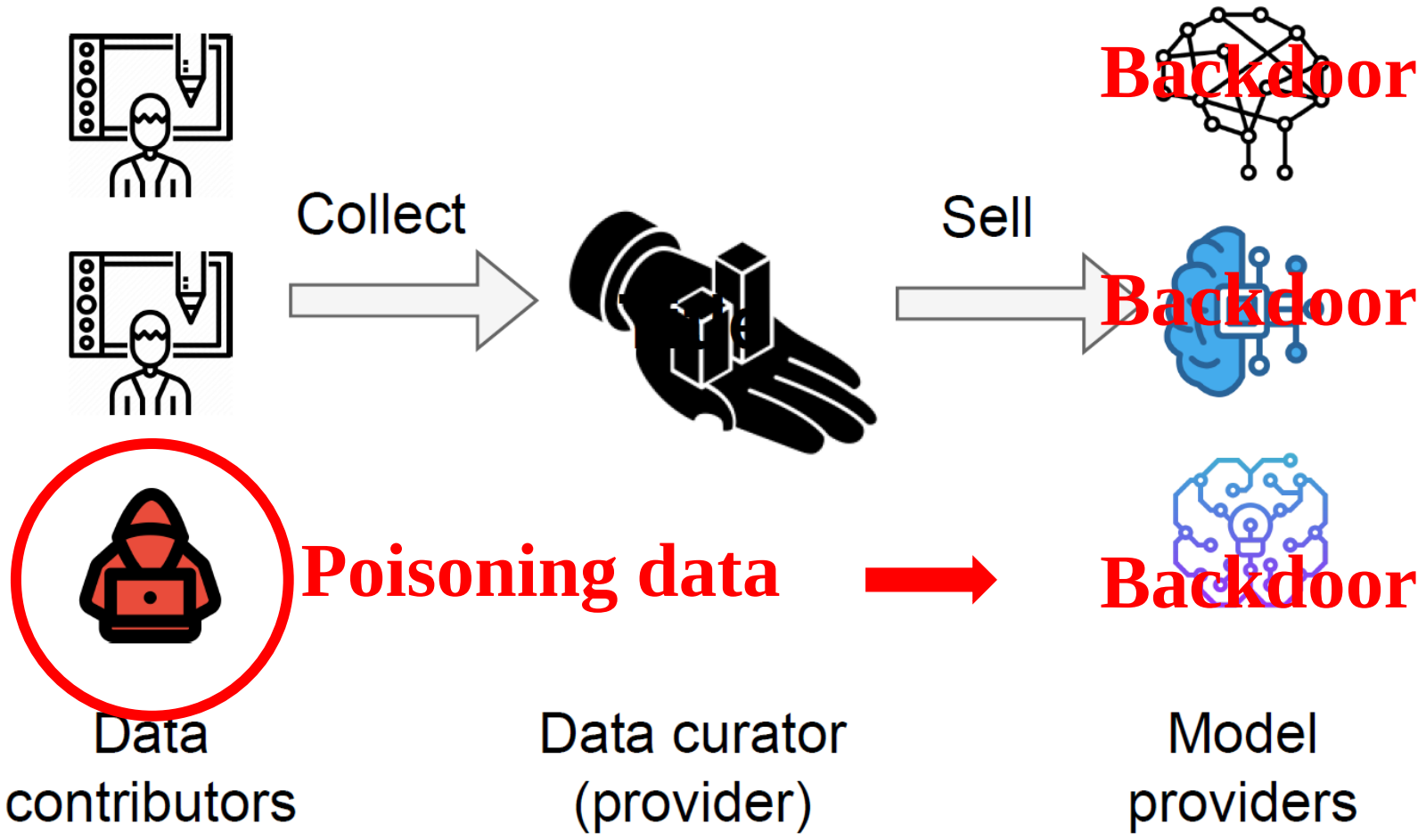
TELUS International

Commercial Data Curators

Data as a Service



Data as a Service



Defenses against Backdoor

- Prevention/removal
- Model based detection
- Data based detection

Defenses against Backdoor

- Prevention/removal

  Indiscriminately applied on underling samples, datasets, or models; often incurs high computational cost and degrades utility. Inapplicable for DaaS.

- Model based detection
- Data based detection

Defenses against Backdoor

- Prevention/removal
- Model based detection

  Impractical for DaaS as different model providers will use different models, each model needs to be assessed.
- Data based detection

Defenses against Backdoor

- Prevention/removal
- Model based detection
- Data based detection

  Inference phase vs Training phase

Defenses against Backdoor

- Prevention/removal
- Model based detection
- Data based detection

  Inference phase vs Training phase

Training phase detection is suitable for DaaS scenario that allows a data curator to perform data cleansing once-off.

Requirements for Data Cleansing

RM1: One-time operation

RM2: Modality agnostic

RM3: Task agnostic

RM4: Poisoning rate agnostic

RM5: Attack method agnostic

RM6: No clean data access

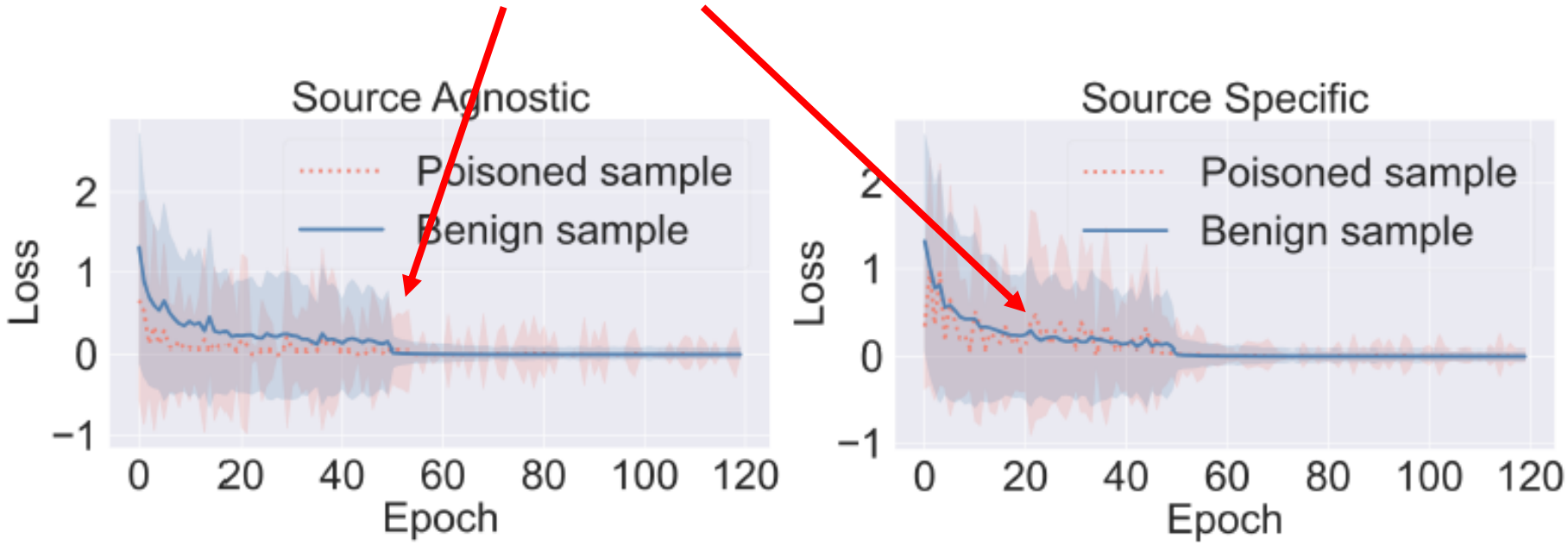
Requirements for Data Cleansing

	Not One-Time Cleansing (RM1)	Clean Data Access (RM2)	Modality Specific (RM3)	Poison Rate Specific (RM4)	Trigger Type Specific (RM5)	Backdoor Type Specific (RM5)	Classif. Task Specific (RM6)
Spectral [31]	○	○	○	●	●	●	●
AC [32]	○	○	○	◖	●	●	●
Spectre [27]	○	●	○	●	●	●	●
SCAn [28]	○	●	○	◖	◖	○	●
Beatrix [29]	○	●	○	◖	◖	○	●
CT [13]	○	●	○	◖	◖	○	●
ASSET [12]	○	●	○	○	○	●	●
Telltale	○	○	○	○	○	○	○

No existing work can satisfy all those practical requirements!

Telltale: Insights

Losses of universal and partial backdoor attack (CIFAR10 + ResNet18)



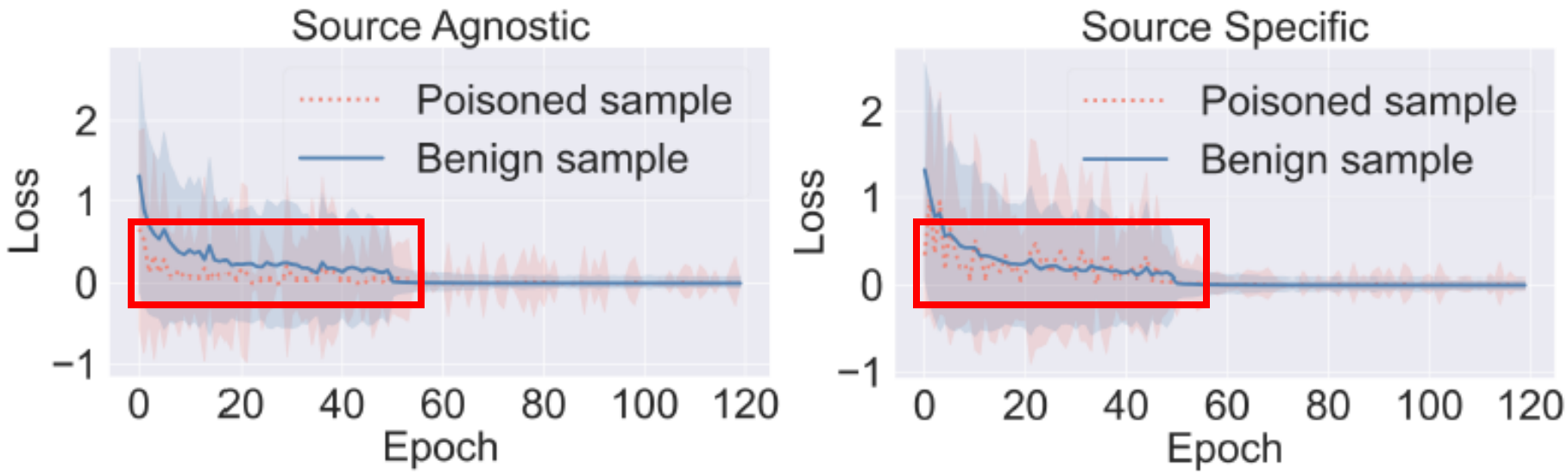
Poisoned sample (dirty-label)

Cat

Airplane

Telltale: Insights

Losses of universal and partial backdoor attack (CIFAR10 + ResNet18)



Poisoned sample (dirty-label)

Cat

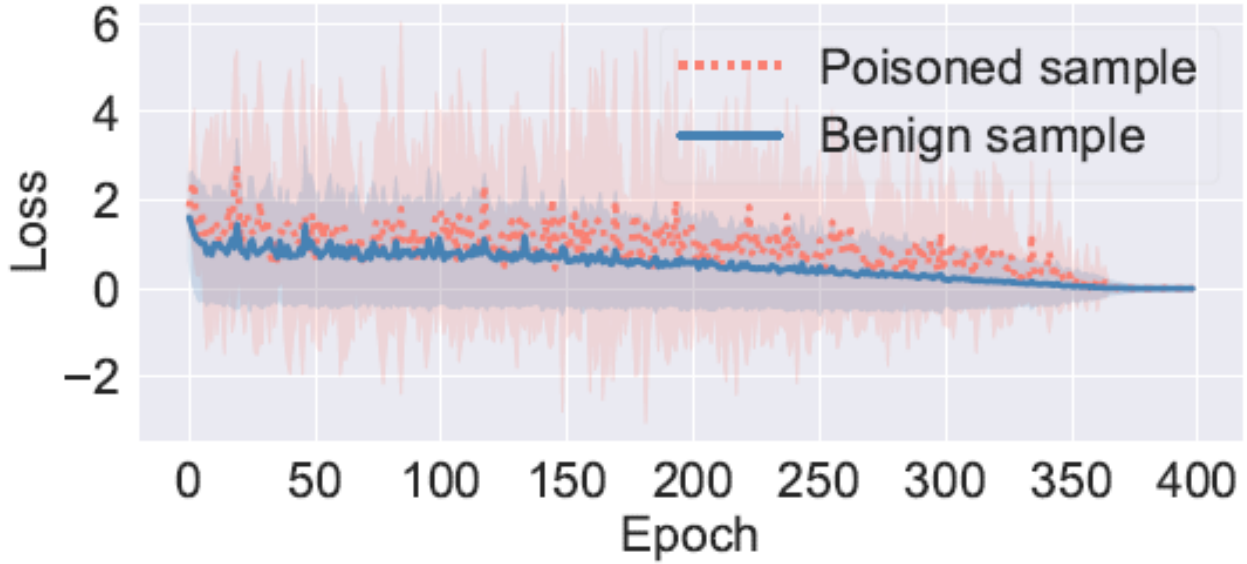
Airplane

Loss of poisoned samples is always lower in early epochs?

Telltale: Insights

Losses of clean-label attack, Narcissus attack CCS'23 (CIFAR10 + ResNet18)

Poisoning rate = 0.05%



Poisoned sample (clean-label)

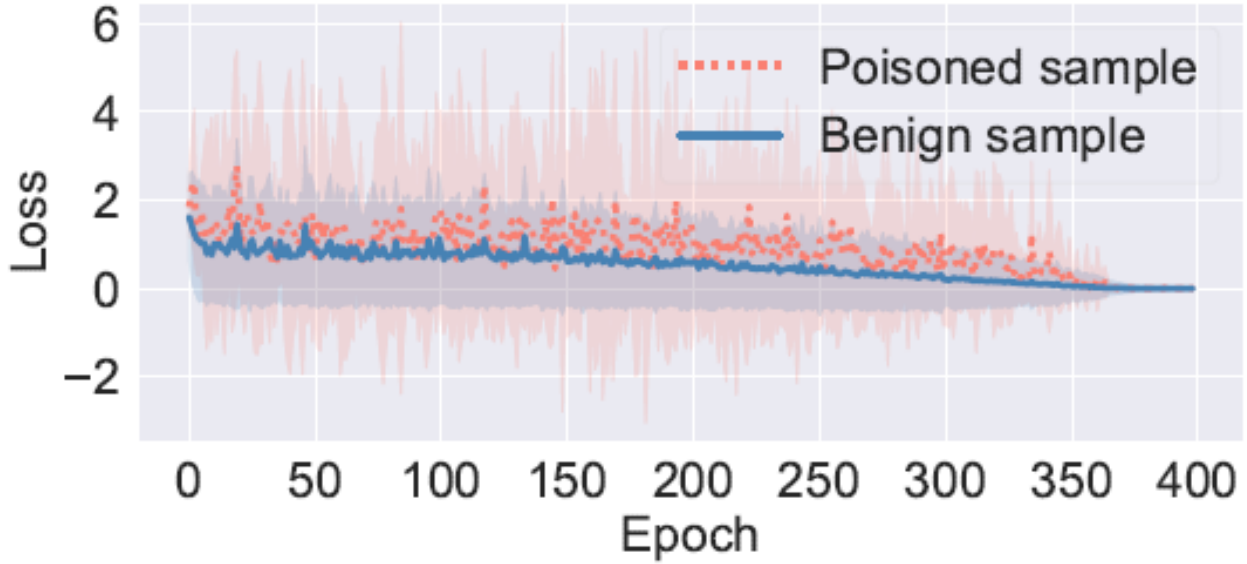
Bird

Bird

Telltale: Insights

Losses of clean-label attack, Narcissus attack CCS'23 (CIFAR10 + ResNet18)

Poisoning rate = 0.05%



Poisoned sample (clean-label)

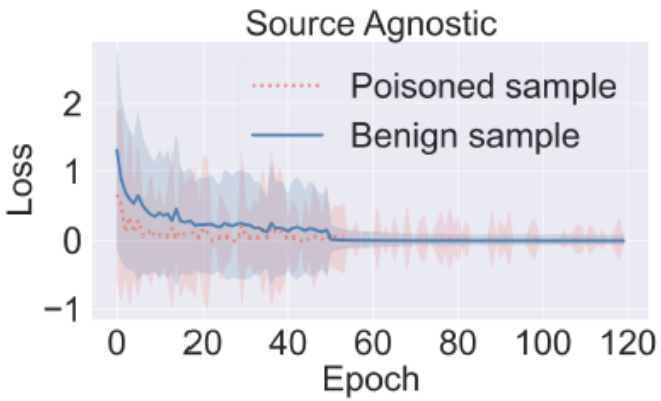
Bird

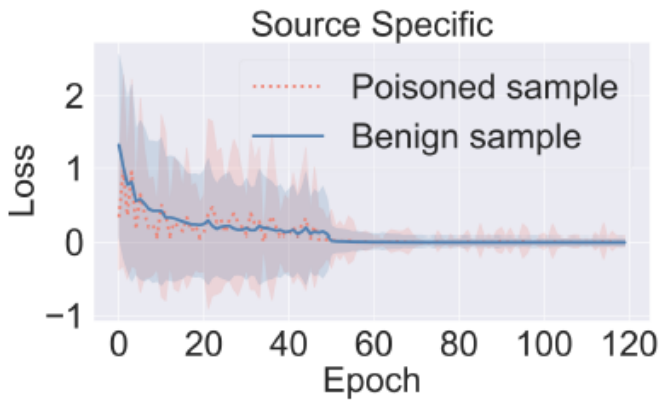
Bird

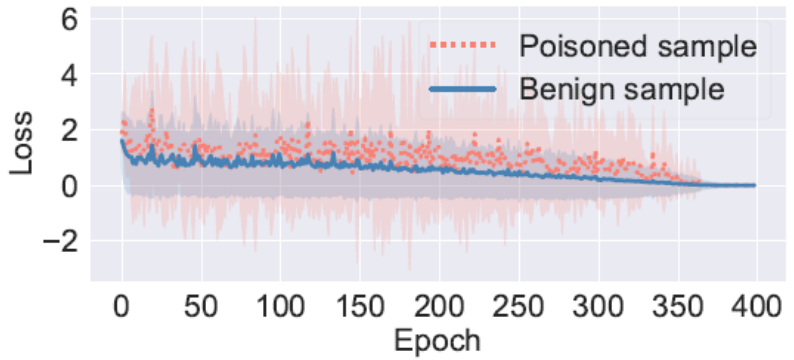
Loss of poisoned samples is always lower in early epochs?

Not really

Telltale: Insights



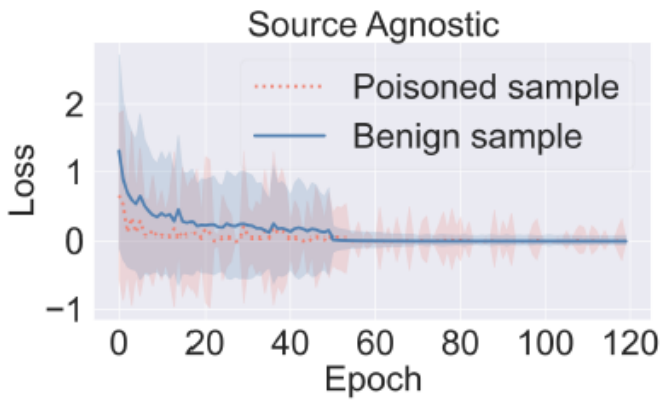


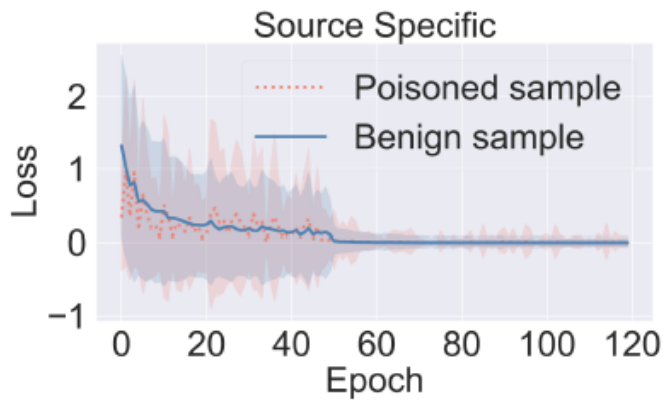


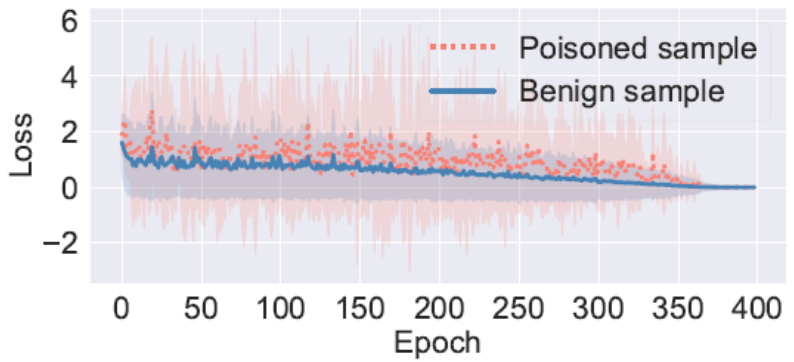
Anyway,
Loss trajectories of benign and poisoned samples are discernable

Satisfy **RM2** (modality agnostic) and **RM3** (task agnostic)

Telltale: Insights



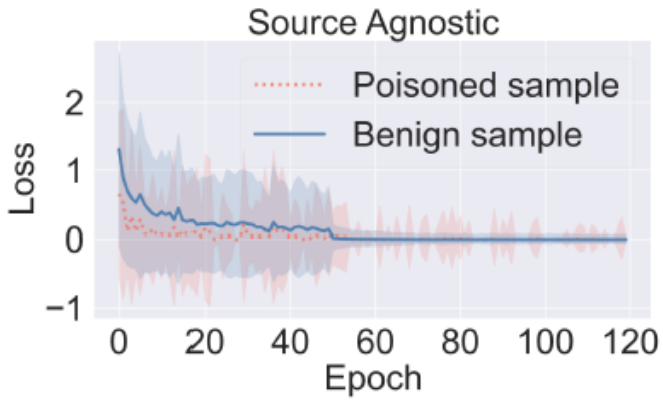


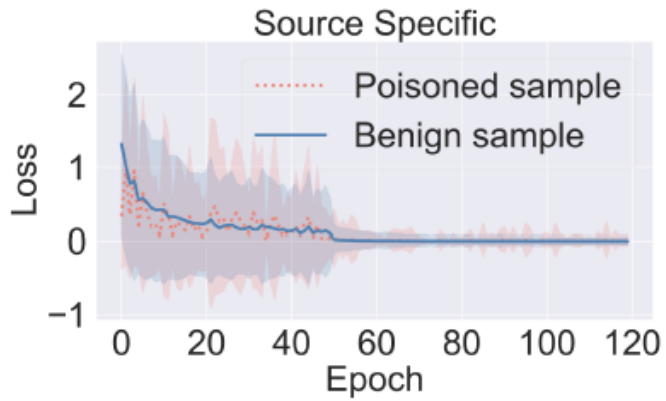


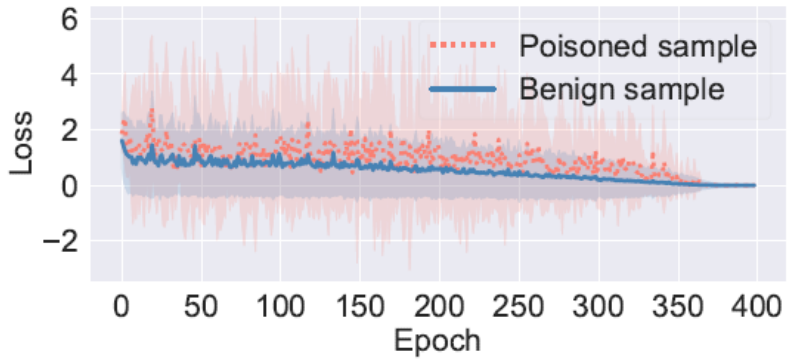
But,

Loss trajectories of benign/poisoned samples are highly entangled

Telltale: Insights





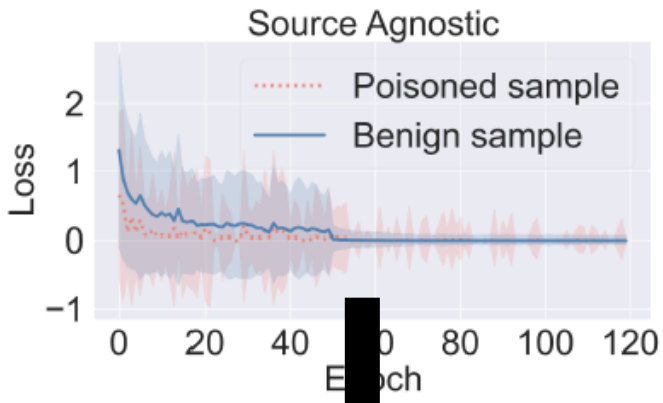


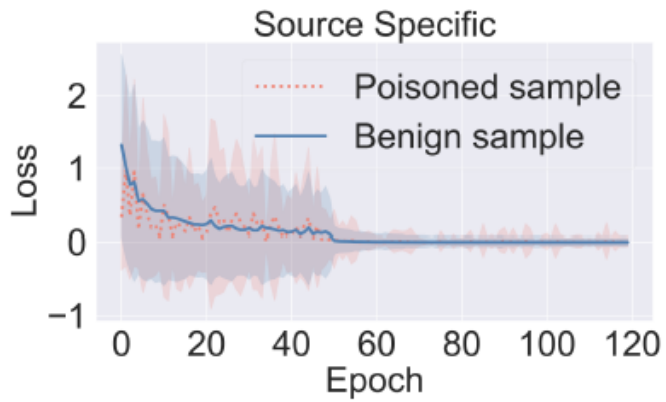
But,

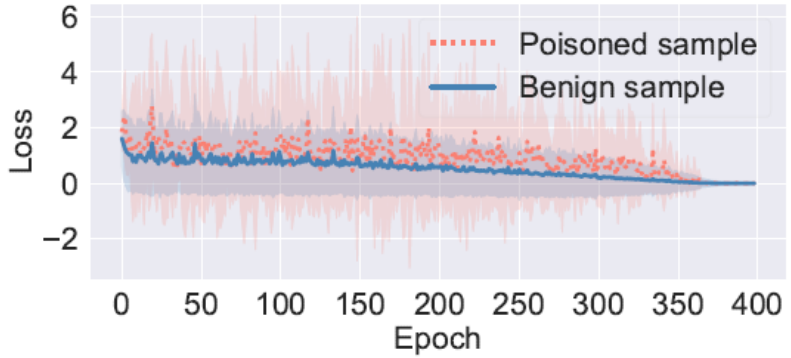
Loss trajectories of benign/poisoned samples are highly entangled

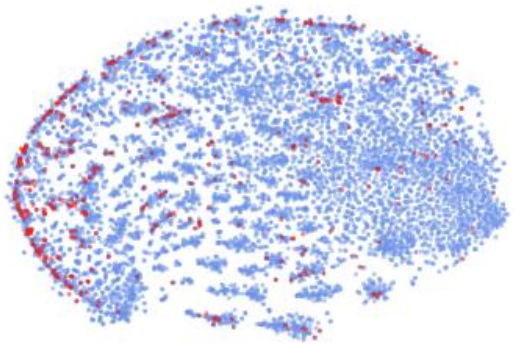
Recall no clean dataset is available for reference

Telltale: Insights





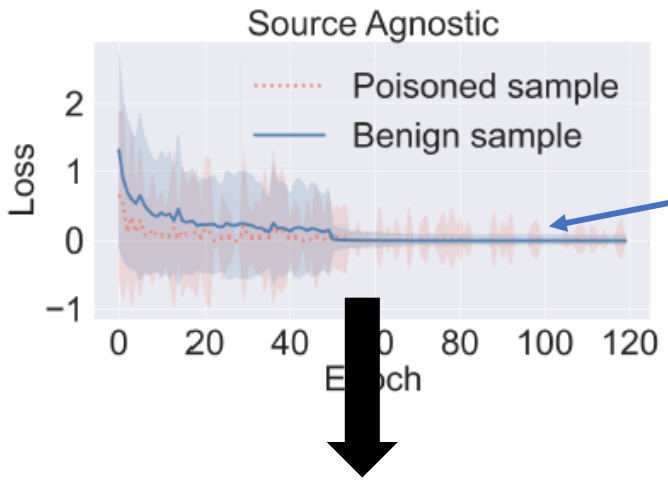




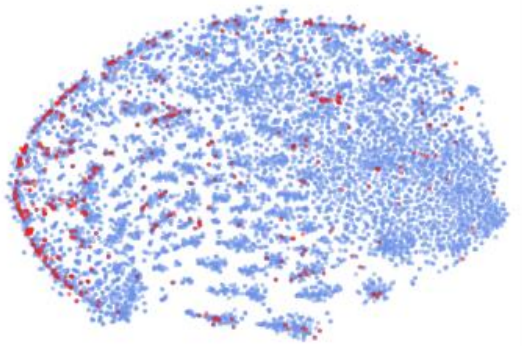
t-SNE

Poisoned samples cannot be separated.

Telltale: Insights

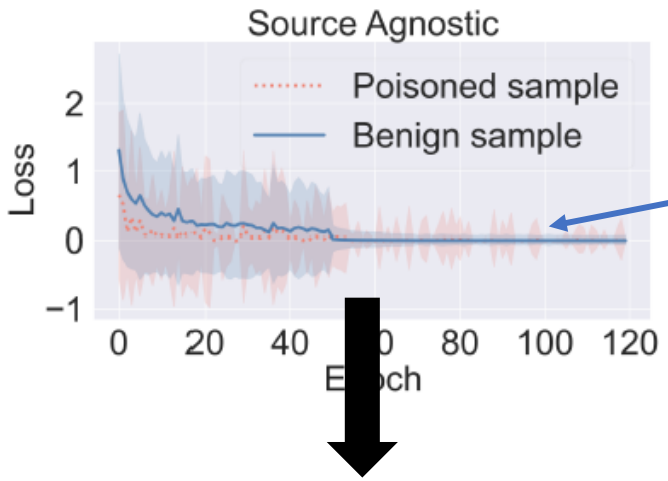


Times-series signal



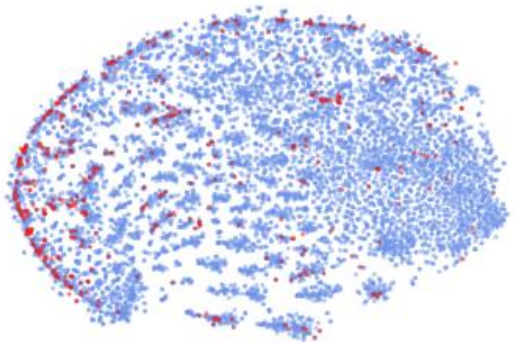
t-SNE

Telltale: Insights



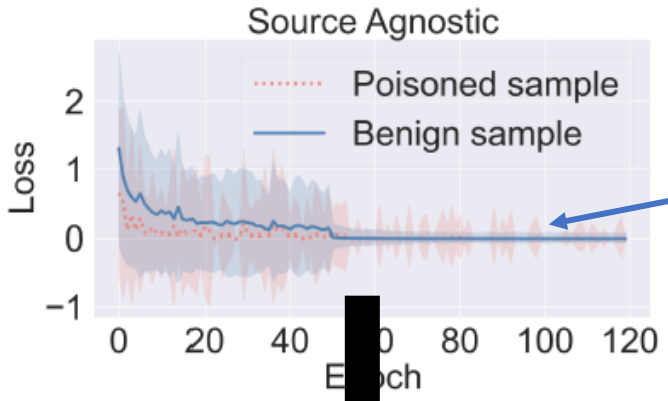
Times-series signal

Spectrum transformation



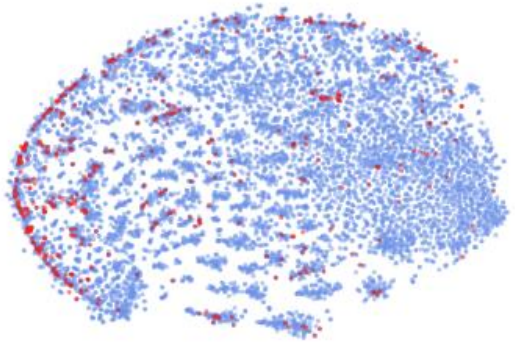
t-SNE

Telltale: Insights

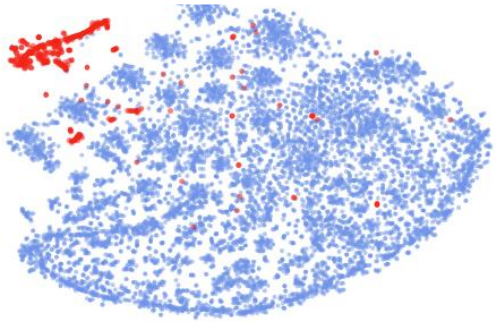


Times-series signal

Spectrum transformation

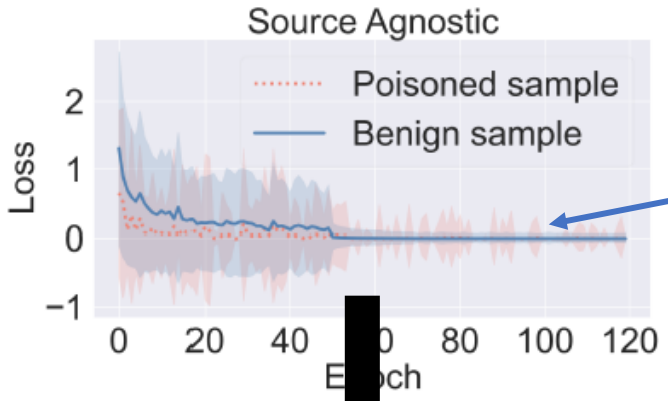


t-SNE



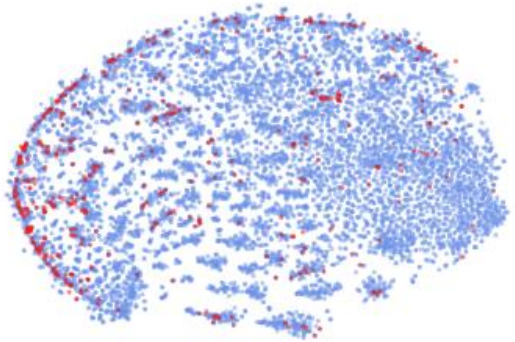
t-SNE

Telltale: Insights

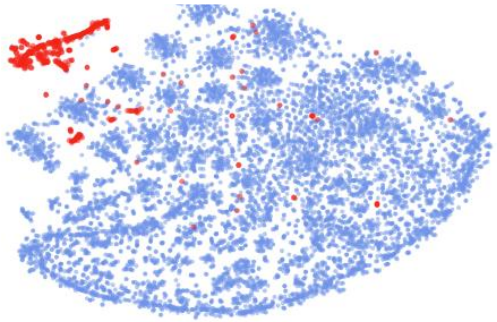


Times-series signal

Spectrum transformation



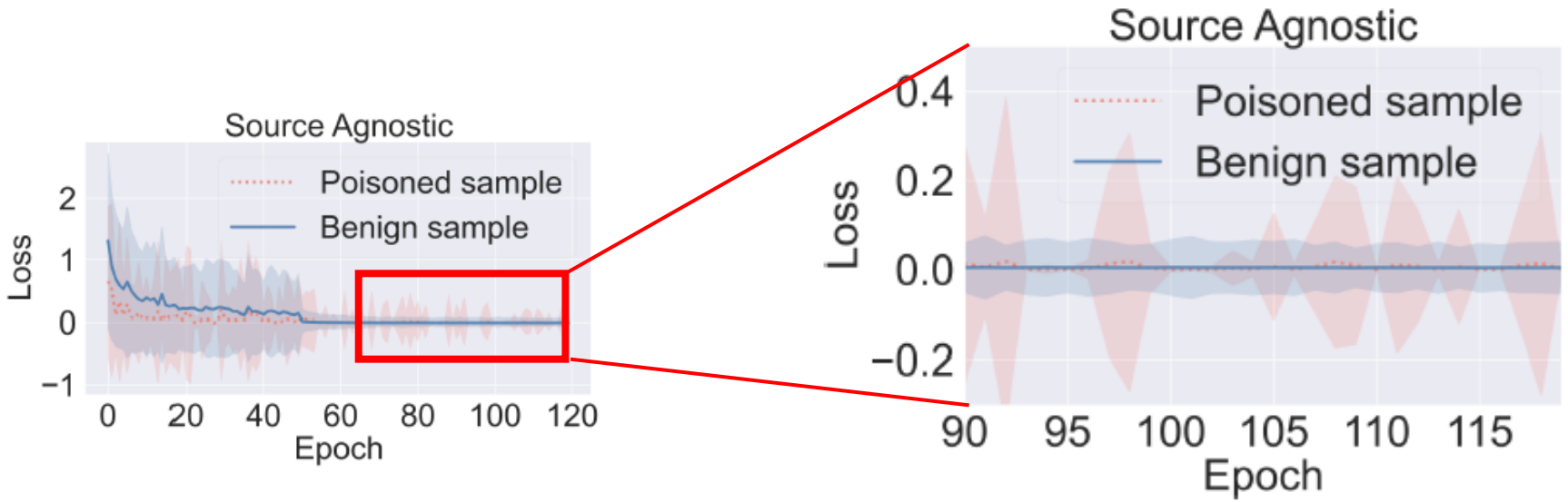
t-SNE



t-SNE

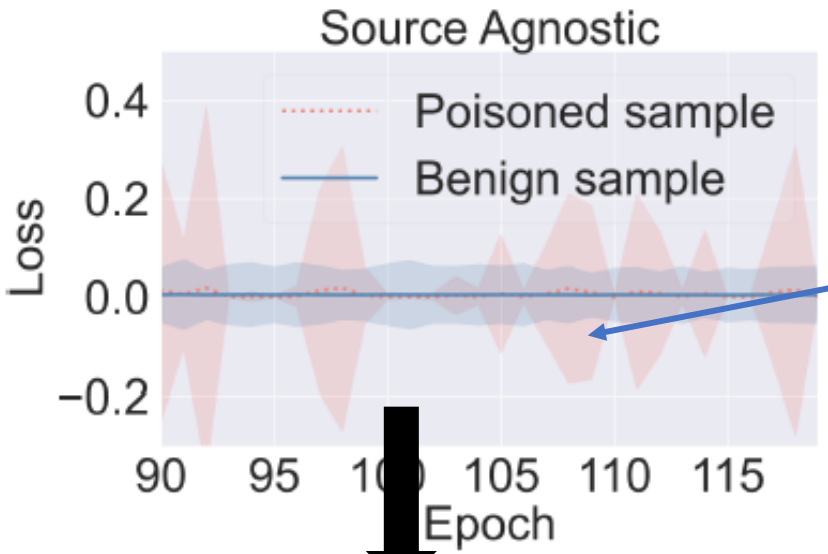
ALMOST SEPERABLE

Telltale: Insights



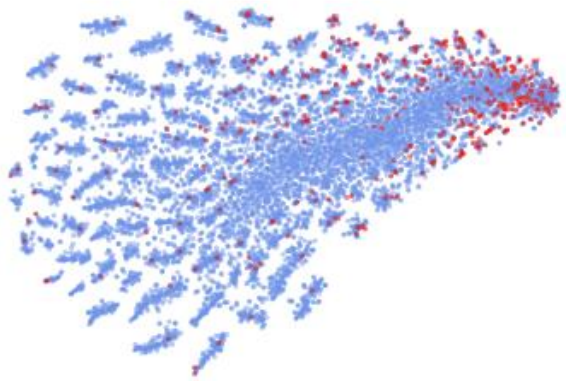
Relative difference is more salient once the model is converged

Telltale: Insights

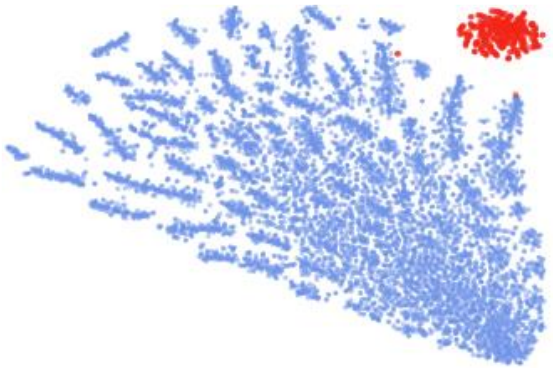


Times-series signal

Spectrum transformation

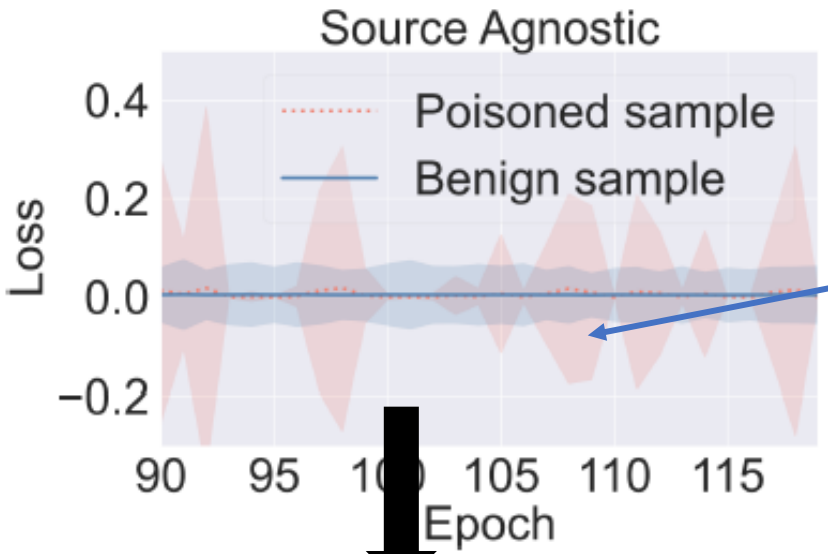


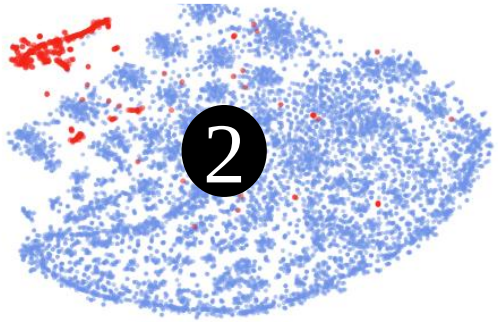
t-SNE



t-SNE

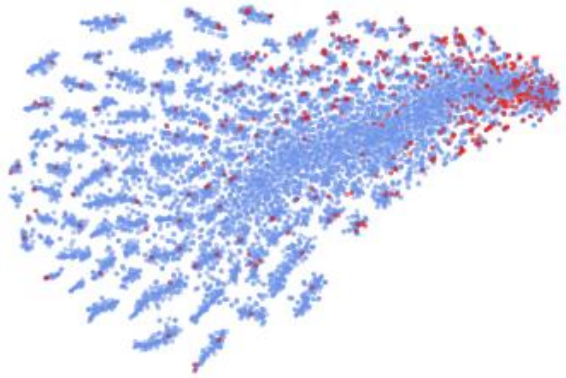
Telltale: Insights



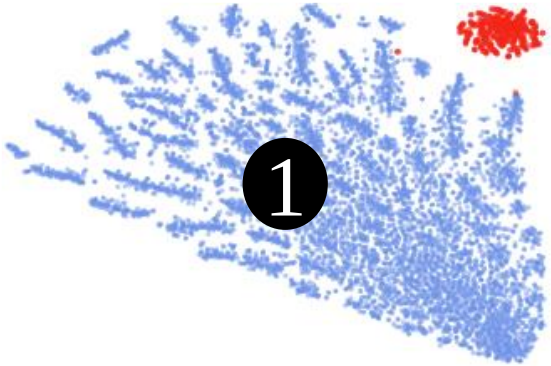


Times-series signal

Spectrum transformation



t-SNE



t-SNE

SEPERATED WELL

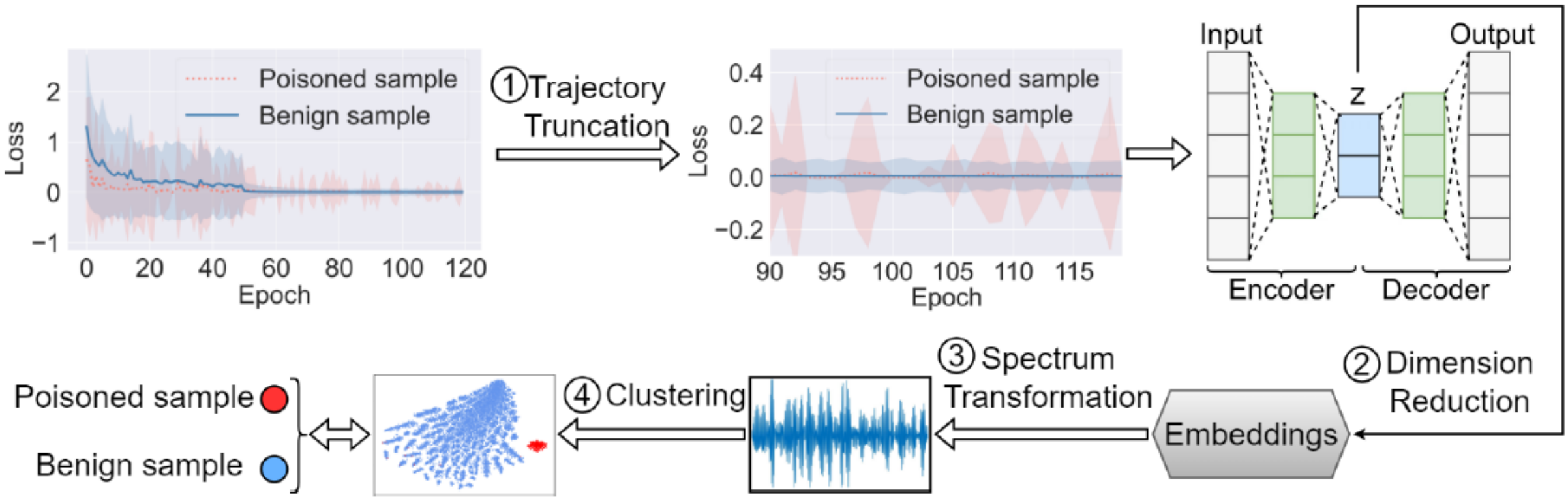
Telltale: Insights Takeaway

Loss trajectory: address RM2 (task agnostic) and RM3 (task agnostic)

Truncation and Spectrum: address RM4 (poisoning rate) and RM5 (attack method)

Clustering: address RM6 (no clean dataset access)

Telltale: Design



DBSCAN is used for clustering because of no prior knowledge of number of clusters (either 2 for poisoned dataset or 1 for benign dataset)

Results: Universal Backdoor

Detection performance against four different triggers (**dirty-label**)
CIFAR10+ResNet18

	Trigger type			
	BadNet	Blend	WaNet	ISSBA
Det. Acc(%)	99.90	99.75	97.32	97.20
FPR(%)	0.17	0.14	0.22	0.23

Results: Universal Backdoor

Detection performance against Narcissus (**clean-label**)
CIFAR10+ResNet18

Det. acc	96.00%
FPR	0.61%

Results: Universal Backdoor

Detection performance at different poisoning rate (**BadNet**)
CIFAR10+ResNet18

	Poisoning rate			
	0.5%	1%	3%	5%
Det. Acc(%)	98.30	99.02	99.13	99.45
FPR(%)	0.22	0.21	0.18	0.17

Results: Partial Backdoor

Detection performance against partial backdoor (**dirty-label**)
CIFAR10+VGG16

Det. acc	97.35%
FPR	0.31%

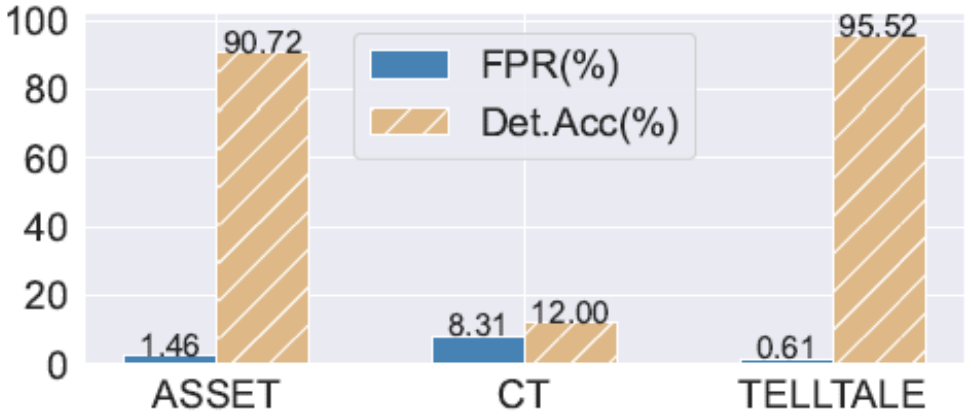
Results: Comparison

Telltale is compared with ASSET (Usenix’23) and CT (Usenix’23) from three scenarios:

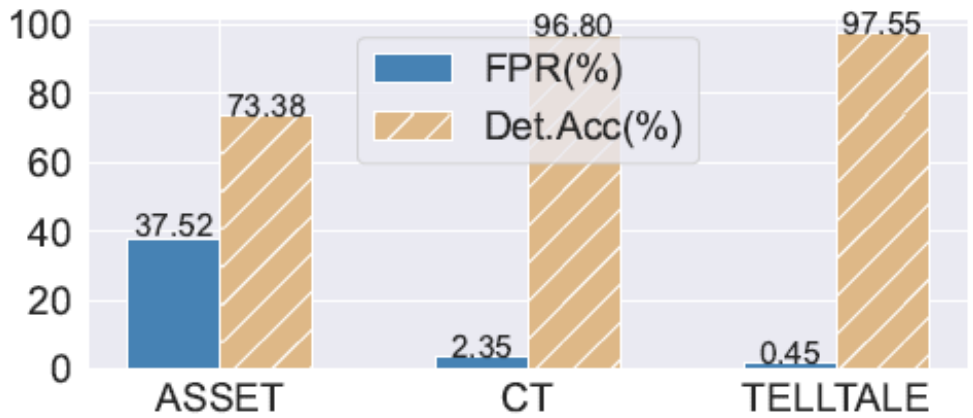
- Narcissus trigger
- Partial backdoor
- Benign dataset

Results: Comparison

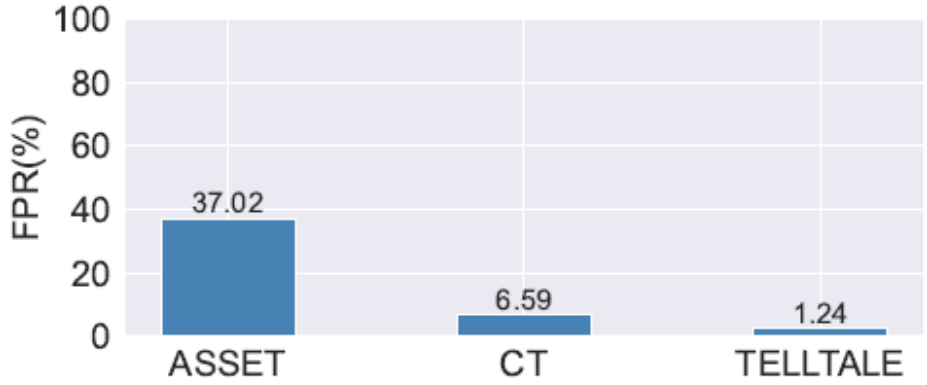
Narcissus trigger



Partial backdoor



Benign dataset

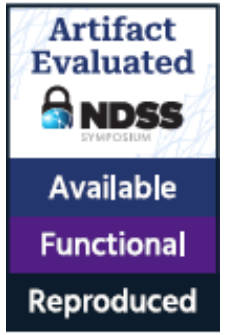


Conclusion and Takeaway

RM1: One-time operation

RM2: Modality agnostic (image, audio, text)

RM3: Task agnostic (classification, regression)



RM4: Poisoning rate agnostic (low to 0.05%)

RM5: Attack agnostic (backdoors, triggers)

RM6: No clean data access


THANK
YOU