CMP-789

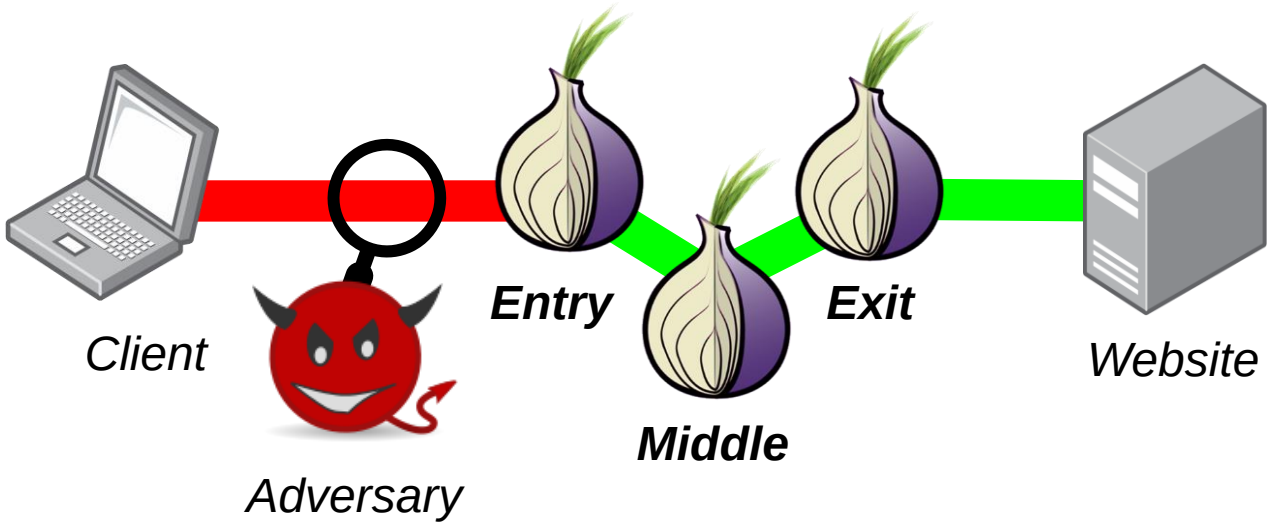
Multi-tab Website Fingerprinting with Deep Learning

May 3rd 2021

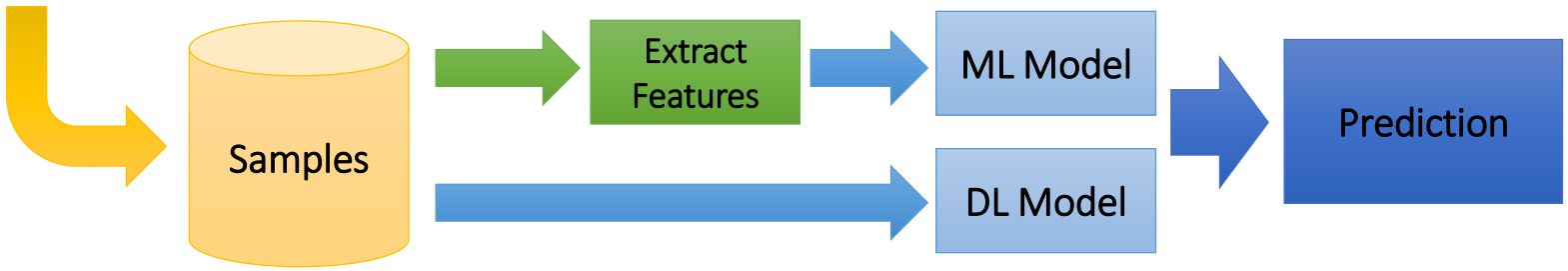
by **Nate Mathews**

Nate Mathews, nate.mathews@mail.rit.edu

Website Fingerprinting (Single-tab)

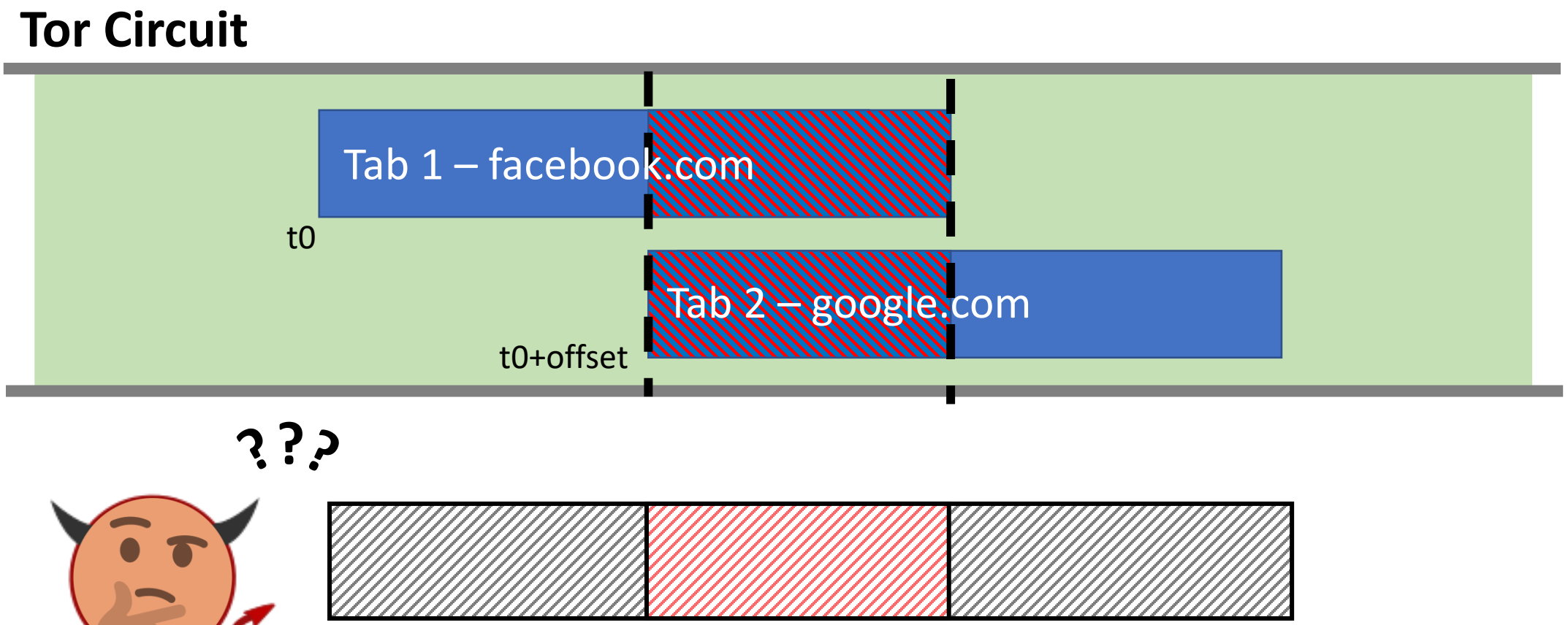


- Use traffic patterns to link client to website
- Up to 98% accuracy w/ 95 sites in closed-world



Nate Mathews, nate.mathews@mail.rit.edu

Website Fingerprinting (Multi-tab)



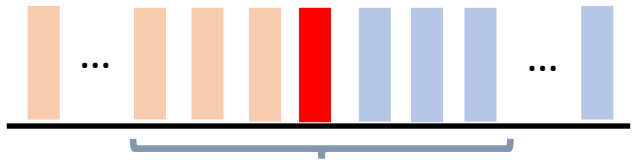
- WF accuracy reduced significantly *[PETS'16]*

Nate Mathews, nate.mathews@mail.rit.edu

WF Multi-tab (background)

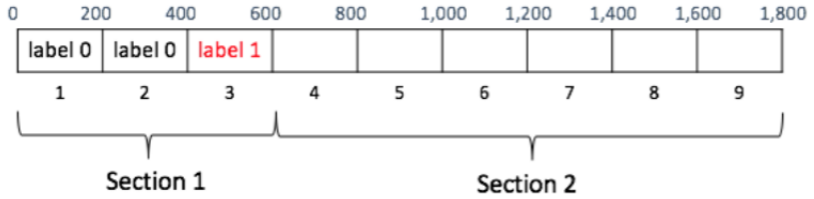
- **2016 [PETS] - On Realistically Attacking Tor with Website Fingerprinting**
  - Generate 23 features for every packet in a sample
  - Use k-Nearest Neighbors to score most probable packet for start of 2nd page
- **2018 [ACSAC] - A Multi-tab Website Fingerprinting Attack**
  - Re-uses [PETS'16]'s 23 features
  - Uses XGBoost with undersampling
- **2019 [AsiaCCS] - Revisiting Assumptions for Website Fingerprinting Attacks**
  - Split sample into blocks
  - Use a Hidden Markov Model to classify each block as *Tab 1* or *Tab 2*

PETS'16 & ACSAC'18



$[f_1, f_2, \ldots, f_{23}]$

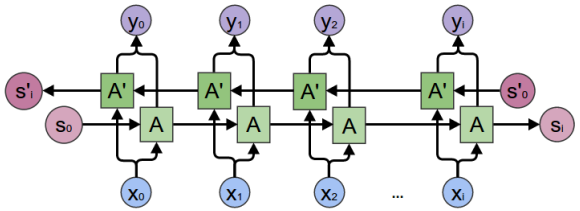
AsiaCCS'19



Nate Mathews, nate.mathews@mail.rit.edu

Deep Learning for Multi-tab

- ***Why?***
  - Prior works only used basic ML and hand-crafted features (see Backup slides)
  - Automatic extraction of features from 'raw' inputs
  - More 'powerful' features learned
    - *DL in Single-page increased acc. and defeated several defenses* [CCS'18]
- ***How?***
  - Treat it like an audio segmentation problem:
    - *E.g. Ingest the sample as time-series data and divide into overlap & non-overlap regions*
    - Bi-directional LSTM w/ convolutional layers for feature extraction [EURASIP'20]



[CCS'18] Sirinam et al. "*Deep Fingerprinting: Undermining Website Fingerprinting Defenses with Deep Learning*"
[EURASIP'20] Gimeno et al. "*Multiclass audio segmentation based on recurrent neural networks for broadcast domain data*"

Nate Mathews, nate.mathews@mail.rit.edu

Hypothesis

H1: Deep learning techniques improve the performance of multi-tab sample splitting when compared to the hand-crafted feature-based techniques from prior works.

H2: Multi-tab Website Fingerprinting attack performance can be shown to be comparable to attack performance in the Single-Tab.

Nate Mathews, nate.mathews@mail.rit.edu

Dataset & Simulation

- ***Dataset***
  - 95 websites with 1k samples each [CCS'18]
    - 50% of samples finish loading within 18 seconds
    - 90% of samples finish loading within 62 seconds
  - Collected using automated Tor-Browser-Bundle
    - PCAP capture begins 1 second before visit request is made
    - PCAP capture ends when the browser receives the "onLoad" event
      - "onLoad" triggers when all static HTML/CSS elements are loaded
      - Dynamic content (e.g. javascript) very often triggers after "onLoad"
        - Ex. Youtube or music player content starts after page load, and so this type of data is *NOT* captured in these samples

[CCS'18] Sirinam et al. "*Deep Fingerprinting: Undermining Website Fingerprinting Defenses with Deep Learning*"

Nate Mathews, nate.mathews@mail.rit.edu

Dataset & Simulation

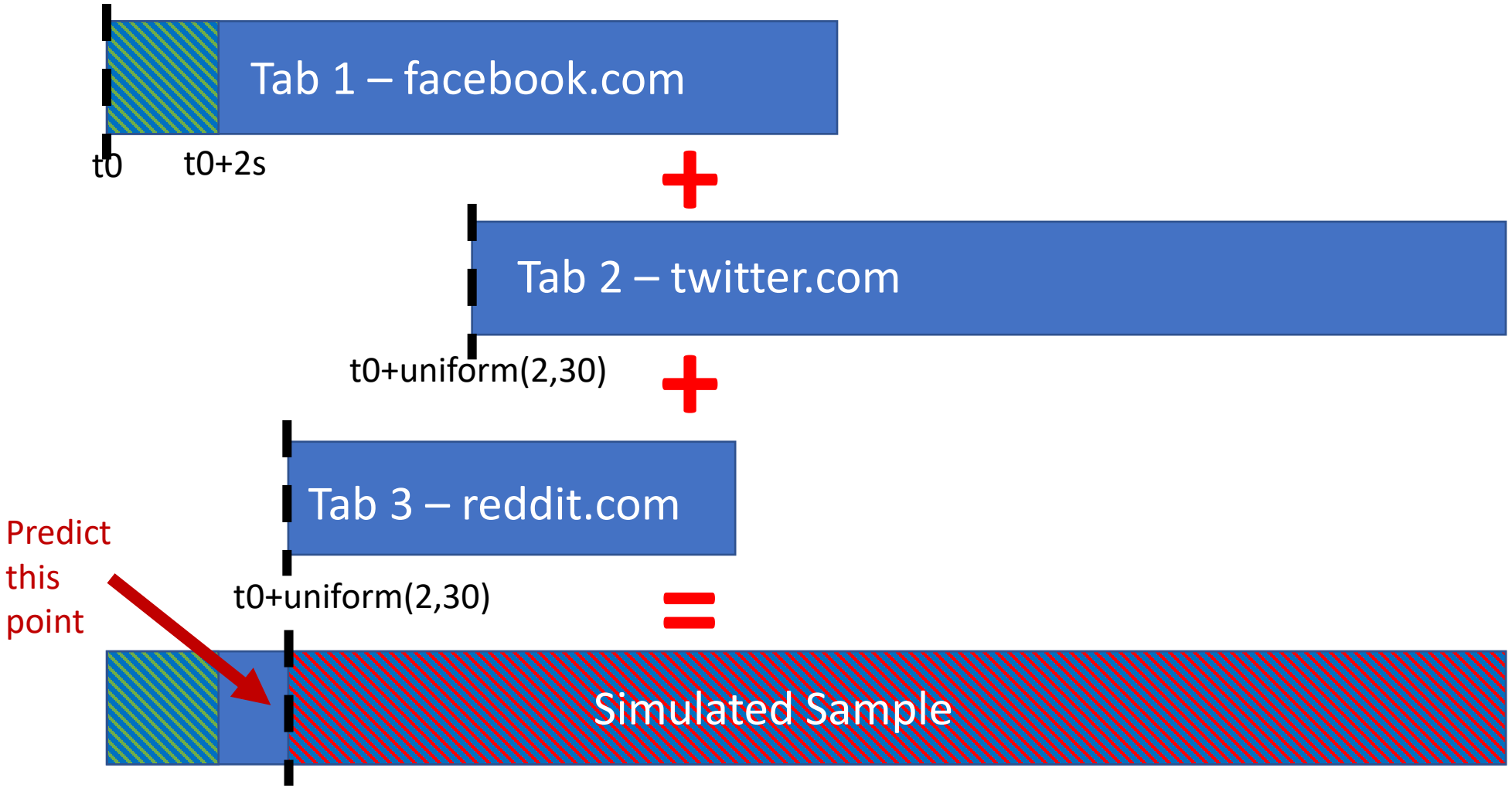
- ***Simulation for Multi-tab*** [AsiaCCS'19]
  - Pick {N} samples at random from the dataset
  - For each {N-1} samples (after the first sample); do
    - Pick a random time offset between *2* and *min(sample_time, 30)*
    - Add offset to each timestamp in sample
    - Concatenate to the first sample, and sort by timestamp
  - Store min(offsets) for split detection

- ***Discuss Limitations at the End***

[AsiaCCS'19] Cui et al. "Revisiting Assumptions for Website Fingerprinting Attacks"

Nate Mathews, nate.mathews@mail.rit.edu

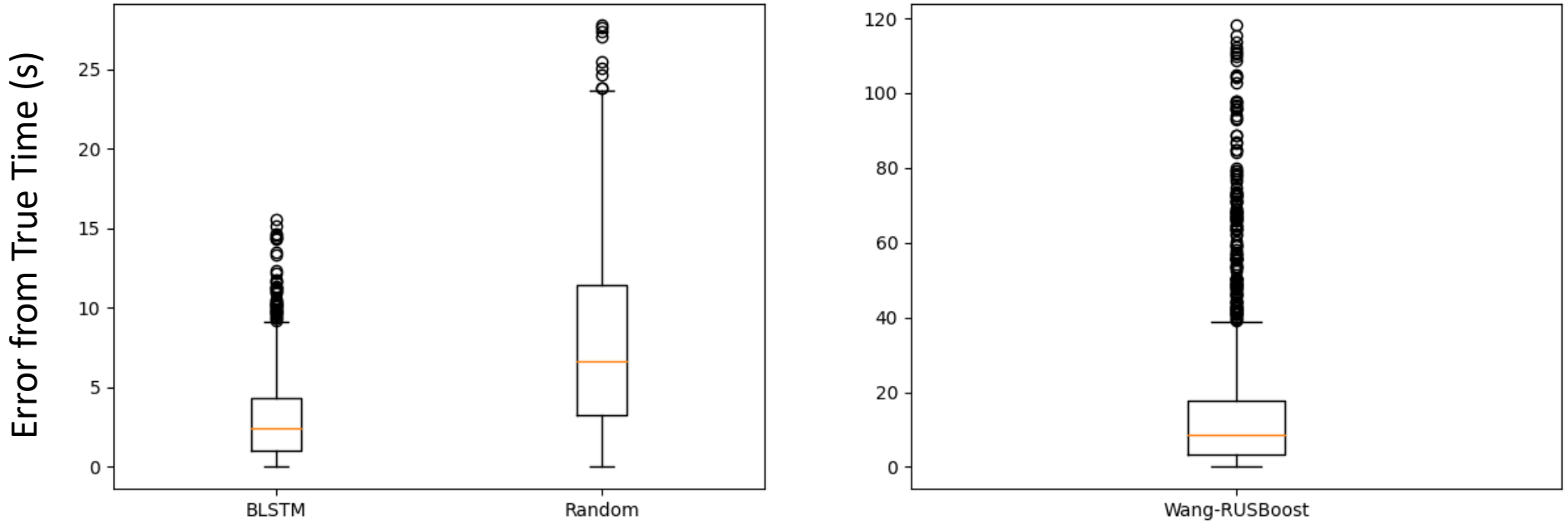
Dataset & Simulation



Nate Mathews, nate.mathews@mail.rit.edu

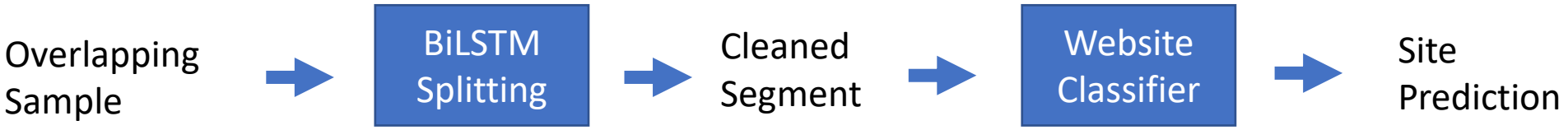
How accurate is LSTM-based splitting?

	CNN-BiLSTM	[1] Features	Random
Accuracy (Counted correct if within 25 packets)	25.2%	14.9%	2.7%



Nate Mathews, nate.mathews@mail.rit.edu

Is this useful? Classifying in the Closed-World



Using CNN Website Classifier from … [CCS'18] Sirinam et al. "*Deep Fingerprinting: Undermining Website Fingerprinting Defenses with Deep Learning*"

Data representation is ***x*** = *time_stamp * direction*

+ Single-Tab ~96% accuracy

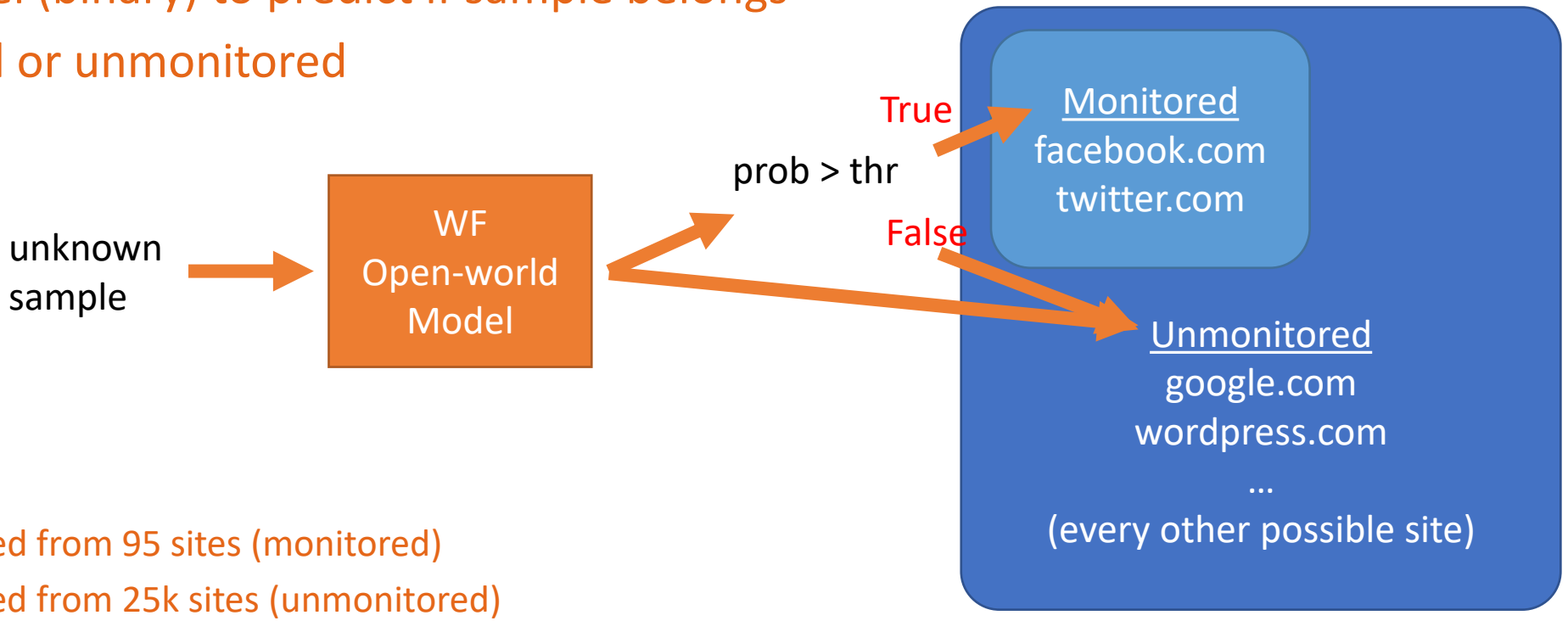
Without Splitting	With Splitting	With (simulated) Perfect Splitting
91.2%	74.8%	91.6%

Perfect Splitting notably out-performs Without Splitting when {N-tabs} > 3
See Backup-Slides

Nate Mathews, nate.mathews@mail.rit.edu

Classifying in the Open-World

- Real-world contains more than just monitored sites
  - Train OW model (binary) to predict if sample belongs to monitored or unmonitored



- OW Multi-tab
  - 2-tab samples
  - 25k samples generated from 95 sites (monitored)
  - 25k samples generated from 25k sites (unmonitored)

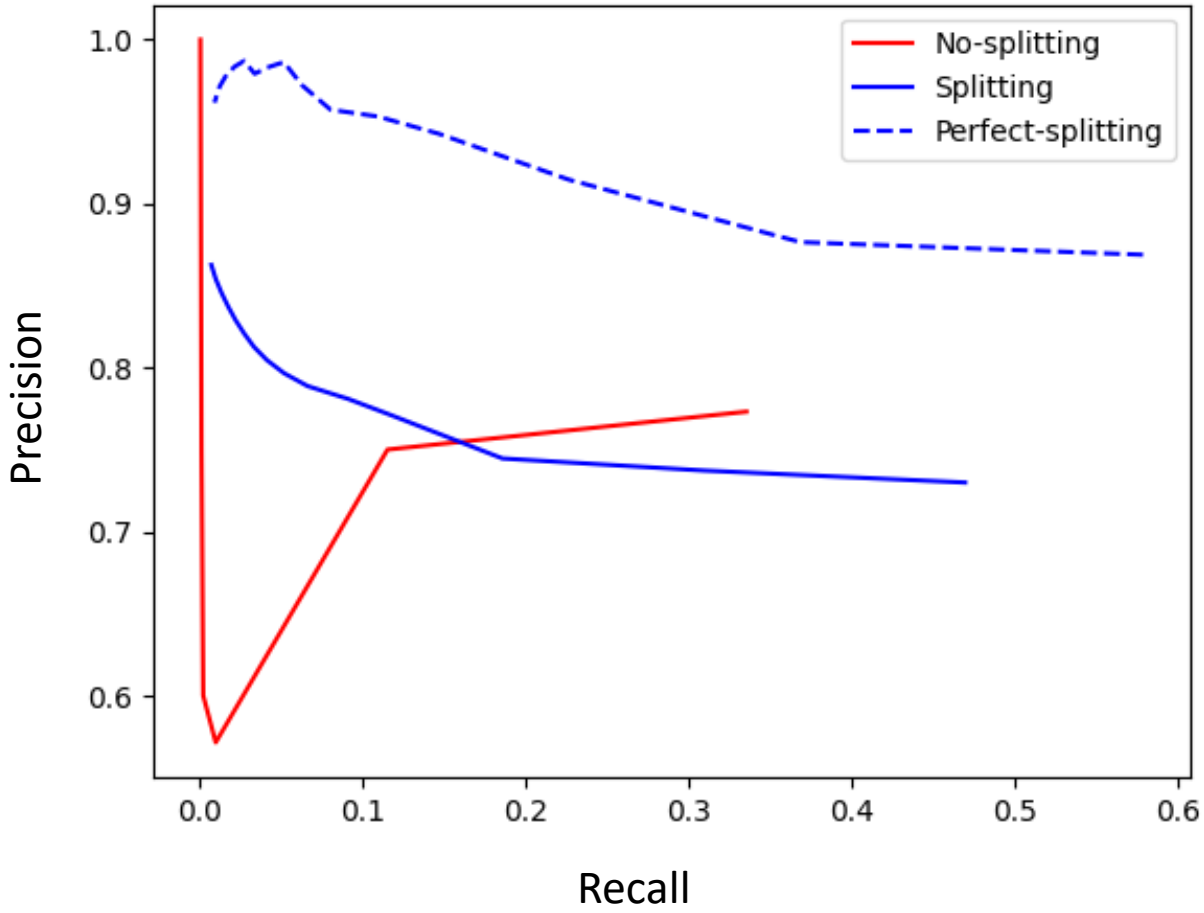
Nate Mathews, nate.mathews@mail.rit.edu

Open-world Evaluation

- Precision-Recall Curve
  - TP = Correct monitored
  - TN = Correct unmonitored
  - FP = Incorrect unmonitored as monitored
  - FN = Incorrect monitored as unmonitored
  - Pre = TP / (TP + FP)
  - Rec = TP / (TP + FN)

- Best Recall

	Prec.	Rec.
No-splitting	77.2%	33.5%
Splitting	73.9%	47.2%
(Perfect) Splitting	**86.8%**	**57.8%**



Nate Mathews, nate.mathews@mail.rit.edu

Conclusions…

H1: Deep learning techniques improve the performance of multi-tab sample splitting when compared to the hand-crafted feature-based techniques from prior works.

Yes. LSTM is notably more accurate than prior ML, but still often has high error.

H2: Multi-tab Website Fingerprinting attack performance can be shown to be comparable to attack performance in the Single-Tab.

No. Perfect splitting is needed in 'hard' scenarios, and LSTM model is very far from achieving the needed performance.

Nate Mathews, nate.mathews@mail.rit.edu

Limitations…

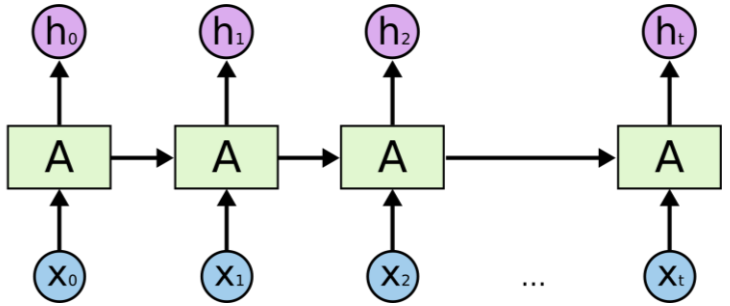
- **Simulation**
  - Samples may be missing additional ‘noise’
    - *Dynamic elements (e.g. Adverts) may produce extra patterns that are not captured in the current dataset*
    - *Multiple page loads may cause congestion (and therefore additional timestamp delays)*
- **Assumptions**
  - Real user-behavior is not known
    - *How often do users open multiple pages simultaneously (and how many tabs are opened)?*
    - *What does the time-offset between the first page and secondary pages look like?*
    - *Are secondary pages often related to the first (e.g. different threads on reddit.com)?*

Backup Slides

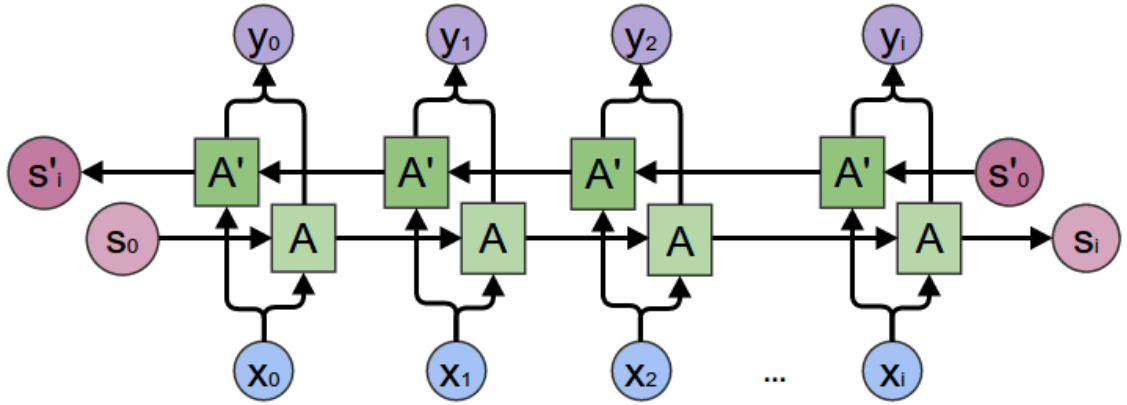
Nate Mathews, nate.mathews@mail.rit.edu

LSTMs Quick Review

Basic LSTM



Bi-Directional LSTM

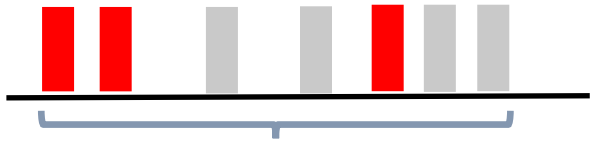


Bi-Directional LSTM allows more patterns to be seen. E.g. The model may find different traffic patterns as the sequence approaches the split point in the forwards direction versus when approached from the backwards direction.

Nate Mathews, nate.mathews@mail.rit.edu

Traffic Representation

- Many ways to represent traffic for DL



directional +1,+1,-1,-1, +1,-1,-1

directional-bursts +2,-2, +1,-2

timestamp 0.0, 0.1, 0.12, 0.21,…

*timestamp * directional* 0.0, 0.1, -0.12, -0.21,…

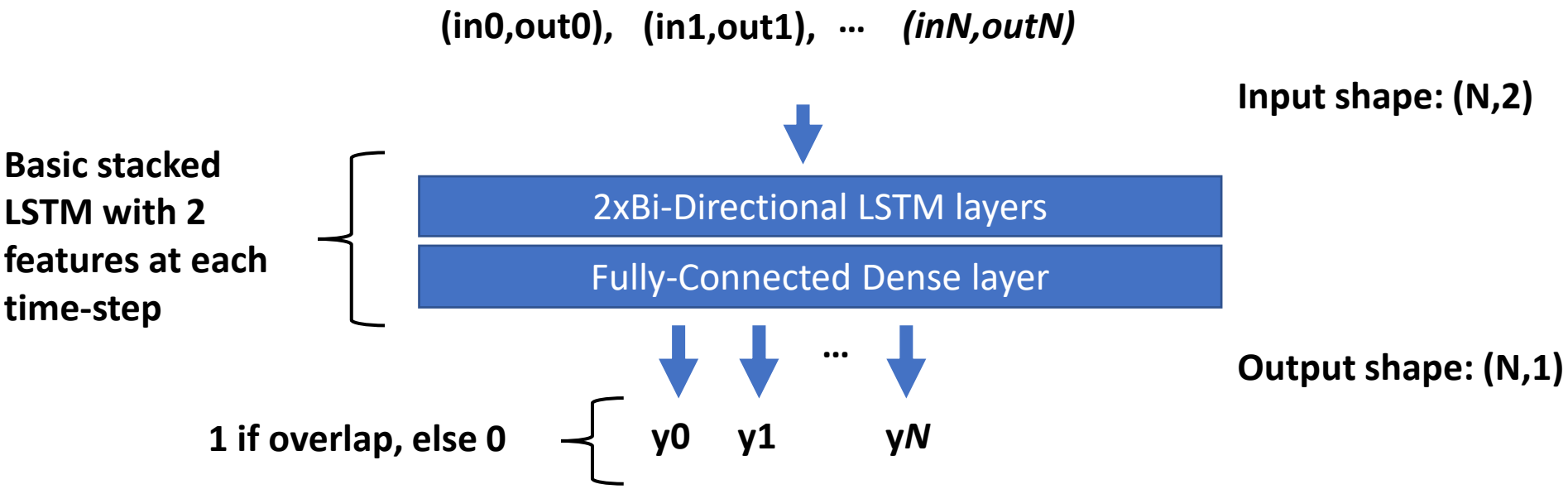
time-bursts +2,-1, +1,-3

These are the representations used with my models.

Nate Mathews, nate.mathews@mail.rit.edu

BiLSTM

X represented using *time-bursts*

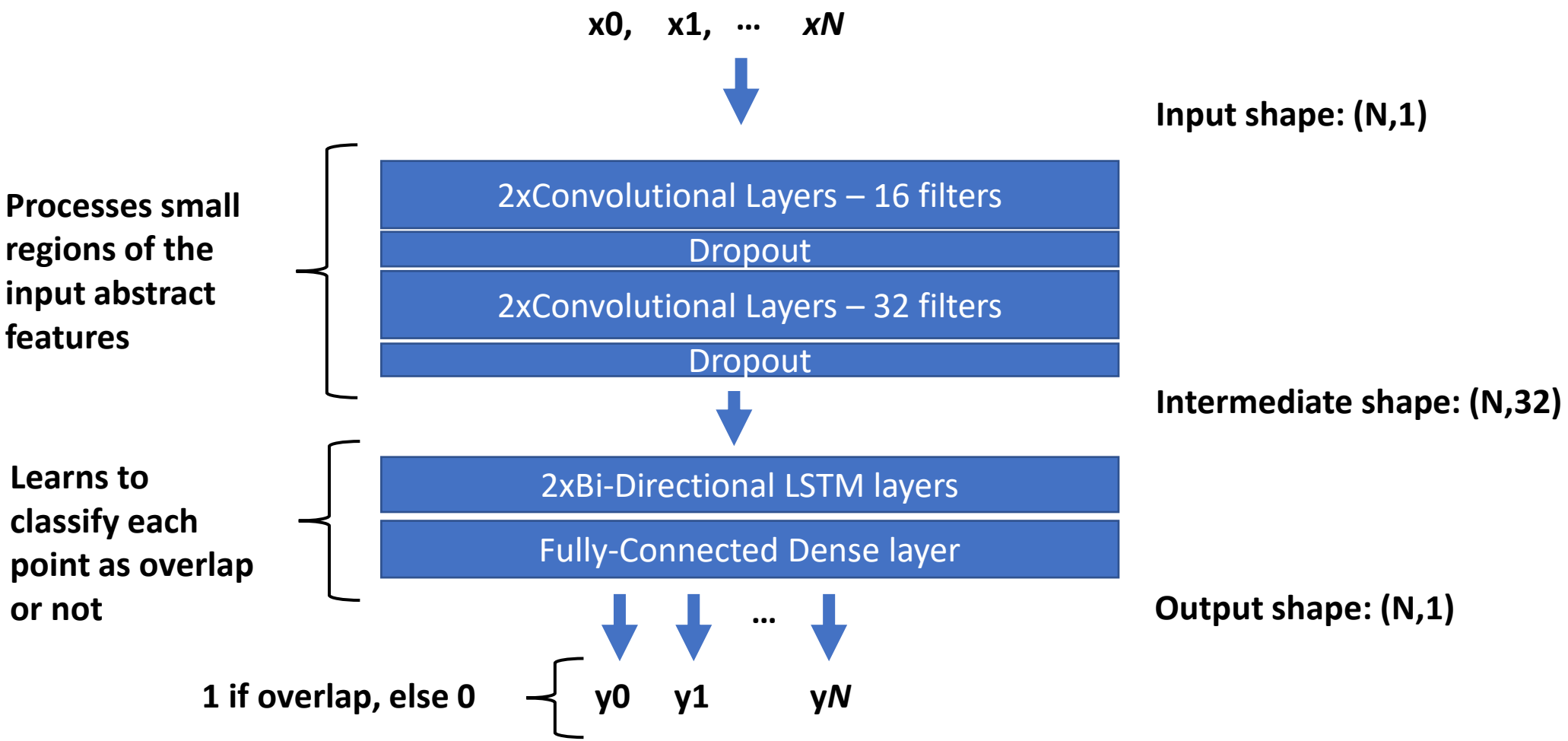


- Cheaper computational cost and easier to train
  - Predictions are coarse-grain (e.g. rather than packet-level, predictions are time-interval level)

Nate Mathews, nate.mathews@mail.rit.edu

CNN-BiLSTM

*X represented using packet_time * packet_direction*

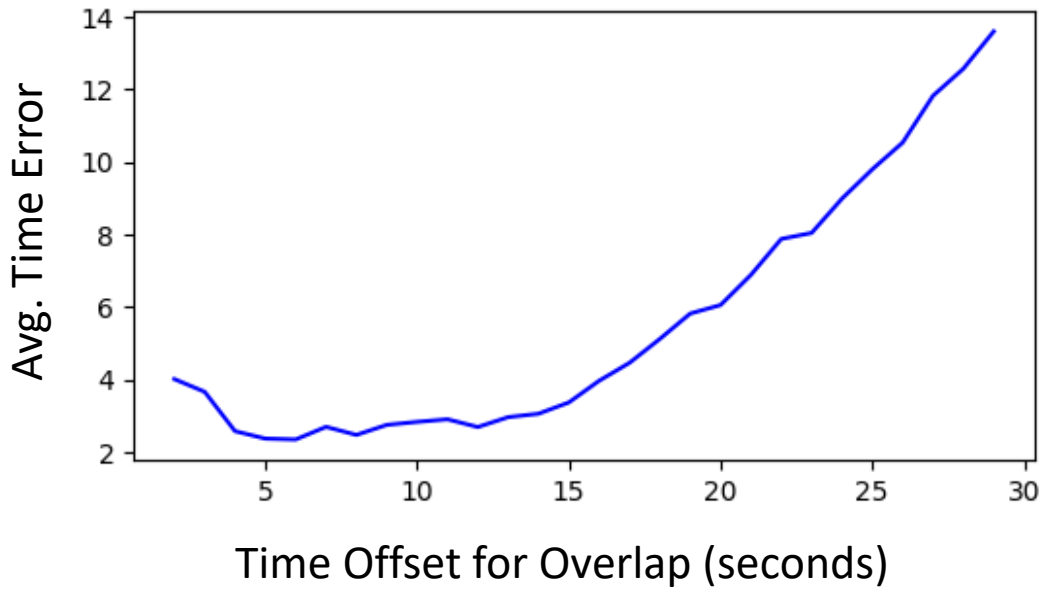


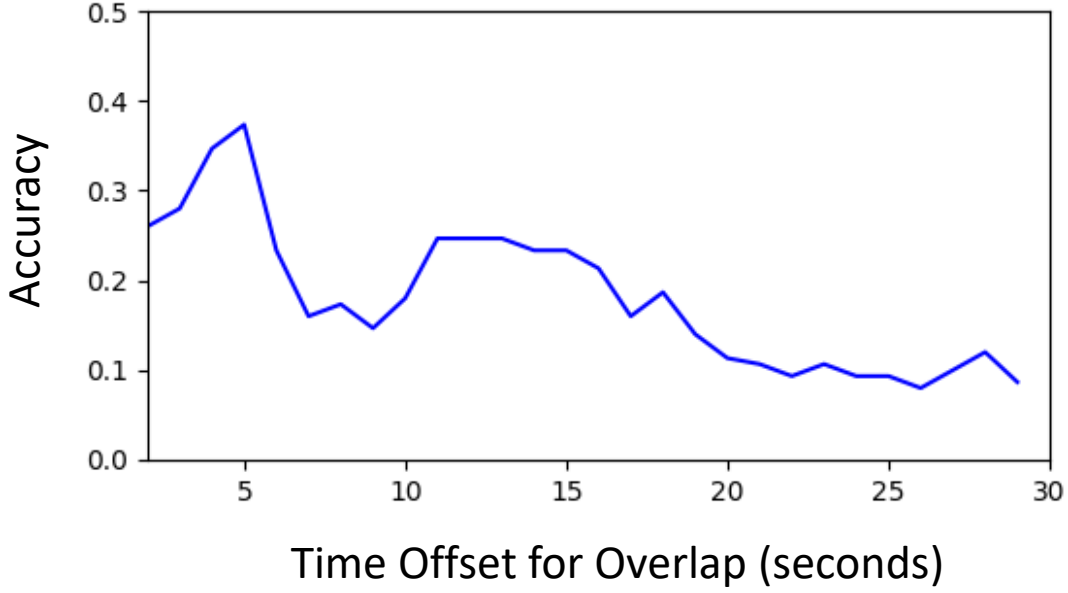
Additional Results & Experiment Details

Nate Mathews, nate.mathews@mail.rit.edu

Sample Splitting Evaluations (2-tabs)

- Performance at different overlap offsets
  - Splitting error generally increases as the time offset for the overlapping sample is increased

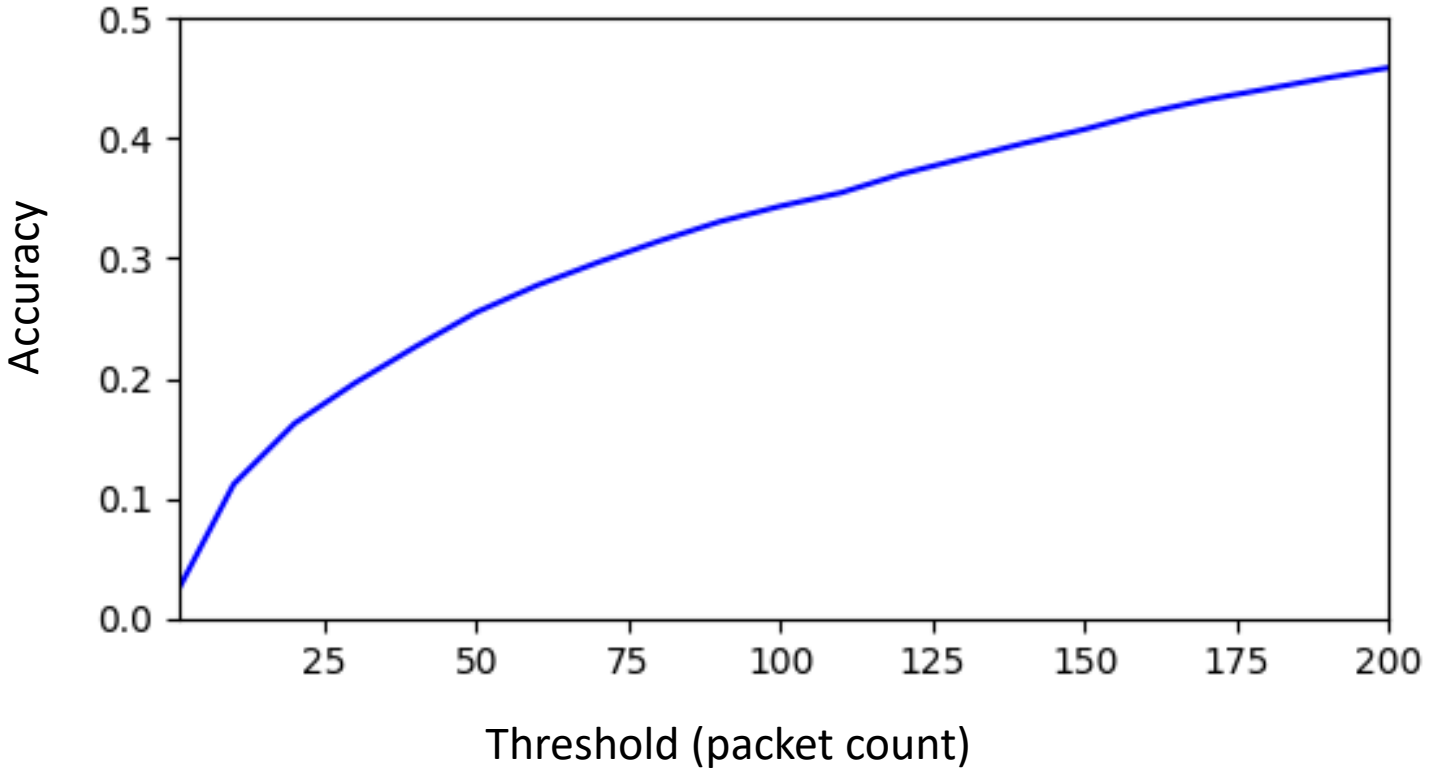




Nate Mathews, nate.mathews@mail.rit.edu

Sample Splitting Evaluations (2-tabs)

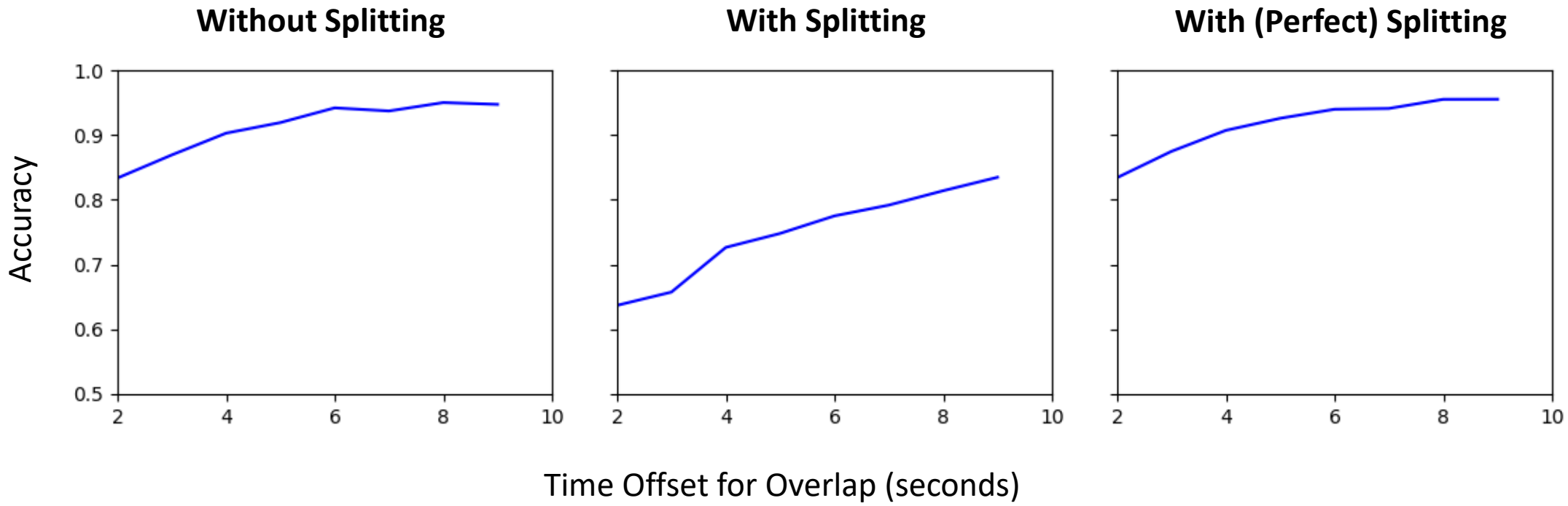
- Splitting accuracy at different packet threshold values
  (prediction counted as correct if within *threshold* packets)



Nate Mathews, nate.mathews@mail.rit.edu

Classifying Split Samples (2-tabs)

- Fingerprinting model performance at different overlap time offsets
  - Larger offsets allow for more clean features to be seen by the WF model, and thus higher fingerprinting accuracy.



Nate Mathews, nate.mathews@mail.rit.edu

Is this useful? Classifying Split Samples

- Performance at with different number of tabs
  - LSTM Splitting almost always performs worse than no-splitting
  - Perfect splitting can be much better when {N} > 3
    - 15% better when {N} is randomly selected for each sample in interval (2,5)

	Without Splitting	With Splitting	With (simulated) Perfect Splitting
2-tabs	91.2%	74.8%	91.6%
3-tabs	69.5%	57.6%	71.7%
4-tabs	52.3%	47.1%	62.0%
5-tabs	31.5%	33.6%	59.7%
Up to 5-tabs	*59.4%*	*48.3%*	*68.4%*

+ Single-Tab ~96% accuracy